

context

Based on [this paper](http://www.plantphysiol.org/content/183/2/637/tab-figures-data#fig-data-additional-files) (<http://www.plantphysiol.org/content/183/2/637/tab-figures-data#fig-data-additional-files>) we want to see in what category our Azolla MYC-likes fall, perhaps reproduce some trees and check if the shared domains coincide with what is described in the paper.

this notebook

In the previous notebook, I confirmed I can align and differentiate the two subfamilies of sequences based on the J&R paper. Here I'm correcting some minor mistakes, and I'm adding some extra sequences of interest.

0 Acquire data

As in step 1

- CHBRA15G00250 Cbrduo1 -> "g8575"
- LC221833 CauDUO1
- LC221832 CleDUO1
- GFZG01001741.1
- Mapoly0017s0071
- XM_024688710.1
- XM_024668389.1
- Sacu_v1.1_s1503.g028048
- Sacu_v1.1_s0147.g023157
- MA_130648g0010
- AmTr_v1.0_scaffold00111.43

And for Subfamily VII

- DN3051_co_g1_i1
- HAOX-0009745
- Mapoly1089s0002
- MN199011
- Sacu_v1.1_s0002.g001222
- Sacu_v1.1_s0041.g012546
- Sacu_v1.1_s0272.g027033
- MA_96853g0010
- MA_10208000g0010
- MA_20462g0010
- AmTr_v1.0_scaffold00038.91
- AmTr_v1.0_scaffold00037.85

Changed in step 2

Added: Azolla MYB like sequences and Arabidopsis GAMYB: MYB33. Azolla sequences were retrieved from fernbase.org; the arabidopsis sequence from uniprot.

Second I'm replacing XM_024688710 with XM_024688730.1 1. This was a typo in the last iteration of this workflow. Also I'm removing GFZG01001741.1_rf_1_GFZG01001741.1_TSA for this sequences behaves oddly in the tree and doesn't quite fit in the alignment. I think it is something different that doesn't belong here. It was an algal sequence so I might be too cautious here...

Now to proceed, let's

1. ~~Acquire data~~
2. linearise and combine
3. align
4. identify domains
5. add Azolla sequences and repeat.
6. for the fun of it, make a phylogenetic tree

I'm using code from my phylogenetics workflow here https://github.com/lauralwd/lauras_phylogeny_wf/blob/master/tree_building_workflow.ipynb
(https://github.com/lauralwd/lauras_phylogeny_wf/blob/master/tree_building_workflow.ipynb)

Clean-up. Since I'm not having version2 or v3 appendices, I'm cleaning up old files to prevent confusion. These may still be retrieved from git history.

```
In [9]: rm data/alignments_raw data/alignments_trimmed data/*_linear.fasta -rf analyses
```

1 linearise and combine

```
In [6]: tree
```

```
.
├── data
│   ├── Azfi-v1-MYB-sequences.fasta
│   ├── MYB33_ARATH.fasta
│   ├── VII_sequences.fasta
│   └── VI_sequences.fasta
├── docs
│   └── step1_differentiate_subfamilies_VI_and_VII.html
├── envs
│   ├── conda-env-jalview.yaml
│   └── conda-env-phylogenetics.yaml
├── figures
│   └── VI-vs-VII_alignment_v1.png
├── step1_differentiate_subfamilies_VI_and_VII.ipynb
└── step2.ipynb
```

4 directories, 10 files

```
In [10]: for i in data/*fasta
do inseq=$(echo $i | cut -d '/' -f 2 | cut -d '.' -f 1)
cat data/$inseq.fasta \
| awk '/^>/ {printf("%s%s\n", (N>0?"\n":""), $0); N++; next;} {printf("%s", $0);} END {printf("\n");}' \
> data/"$inseq"_linear.fasta
done
```

```
In [11]: #cat data/*_sequences_linear.fasta > data/combi_sequences_linear.fasta
cat data/VII_sequences_linear.fasta | sed 's/>/>VII_/' > data/combi_sequences_linear.fasta
cat data/MYB33_ARATH_linear.fasta data/Azfi-v1-MYB-sequences_linear.fasta >> data/combi_sequences_linear.fasta
cat data/VI_sequences_linear.fasta | sed 's/>/>VI_/' >> data/combi_sequences_linear.fasta
head data/combi_sequences_linear.fasta
```

>VII_HAOX-0009745

GLKKGPTAEEDVILSEYVMKHGEGNWNLIQKNTGLPRCGKSCRLRWANHLRPNLKKGAFSREEEALVIKLHAEIGNKWARMALQLPGRTDNEIKNFWNTRIKRRIRAGLPLHST
DLVLCPIATTTTFREKLTEYMEESRDTKPIDRSDDCDGHNTSAHVKESSQT

>VII_Mp3g05910.1

MELGSAPDFSEDVGALKKGPWTS AEDAILVAYVTKHGEGNWNNSVQKHSGLYRCGKSCRLRWANHLRPNLKKGAFTPEEERMIIE LHAKLGNKWARMAAQLPGRTDNEIKNYWNTR
IKRRMRAGLPVYPPEMQNPAANSQYLFEHGEMSMSSGGESECDPGSSPSTGDFQNLVSGMHGGCRMKSCSSLTGMSDPLSSVVTQTL SAGQSISSPGRRMKRMHRDTQCSSMS
GASGGGGLFPQLSDESSKTMYPFKARRSCGTRNSMRLAQIAGFPYDPDPEGLHFEGMPNHGYLNLPPFSCSRPNSSLKLELPSSQSAESADSAGTPGSAMTSPYTA FSLPQNHHL
LSSEADSFSGNSNSNSSFLQALLQEAQNLDPRDQIRSAELSDQLLVLT SANPPMDVSALMSPRKSRWGEDSDPTTPLEGRTYSMFSEDTS PNCTSNWDETSTLQSPLTTVSSNL
QAAHVGGKLIENSAQHDMPCGNYGDEENLISSLLDFARPDATPVVEWYNPPDVYTLGGQPCHSLPEAIEAAFH HQDVVAELEHLGAAGHPVANHVWELGSCPWNNMPGVCQLGDL
PTDCRPLTSIADQMNDPCAIC

>VII_ENA|MN199011|MN199011.1 Selaginella moellendorffii

MGDPMQGGVAAAAALELCEEGRGRRGGGGAVKGLKKGPTPSEDAILVAYVHKHGEGNWNVQKNCGLSRCGKSCRLRWANHLRPNLKKGAFTPEEERTIIE LHAKLGNKWARMA
SQLPGRTDNEIKNYWNTRIKRRMRAGLPVYPDLQDLC SNLARGGHRKEAQDDHYLGHHHNQVVSSSTSSKSSNSNNSSRKSGGILIAAAKNSHDHPSSIATSSYYNDGARG
DDHHDDFAAAYHHHRLLEQINQQHQQQSQPESTSESYGSNSGGSGGNFLRDVLFHQDHQAQRDHDHSPDEQLVYGKSSAANE EGIYQLRLFVWDEDQVEQTTTATTFRGGL
VYPDEDFYALLELESQMPGPPPELIPVPVNLLSYSSGLNHPANLALMFQGE MIPALNSPSTTQLNCYPCLYNRDELPDLYQQQQQQLMWD

>VII_Sacu_v1.1_s0002.g001222 Myb transcription factor [0.077]

MENRKVWDYEGRSTPKKSGKEKHDEGGRVQRKEHGLAAAEVMKKGPTAEEDALLAYVSKHGEGNWNNSVQKNAGVMRCGKSCRLRWTNQLRPGLKKGSLTPQEERLVIEQHALL
GNRWARIAAMLPGRTDNEIKNFWNTRMKRNL RAGKPLYPADVKLVKPAEPTVSNYDAAADWRRQEEERGCGSAAATDREAARHDMIHHPVETAGHALNSVQSRSRSTNFSFL
DMQPPGDL SFGMQFHSYYHSSKRPSYNAQIPAAIDSPSYFYSDHYDESSMFAHLFPNILEVQRGP EEHGLVPNGTFACAAACDNKDIMGANFNDCSSSSSITSTRYPFYRSAYN
GSPESCLACHDAELPSVQMAESADSSSGLSSSASPFVFCNAVPLSEADSF GATNKESPIHKRNGNLVDVLHMIKQEV DASAAGGGDNIIVEDLDLELLVDTSCNASPTSSSNYKS
SVLSANNPLTLLGLTSVNDGDDFGTLLISEEDSELNLHTSRDTRDFAYQFPAAAGTELSHLKLKQQQADTIKSEGPSVSQYVDEELLTLLMLEKPDQGLPTVEFCNENANPVRPT
VNGESQEMIMSGQGELEAMLRYVYTQSDAAEQVMIGSVTVGWGDGSFTW DKSLEIMDEFPSVVTDYVSPAGCSKISP

>VII_Sacu_v1.1_s0041.g012546 Myb transcription factor [0.077]

MERRLSSISHSAYLRADDSVHTFQDMKEQEHL DHYLQFEFSEDQATPLKKGPTAEEDALLAYVSRHGDGNWNTVQKYS GVFVRNGKSCRLRWTNHLRPNLKKGAFSPEEERII
IEQHAAIGNRWSRIAAMLPGRTDNEIKNFWNTRKKRRSRAGLPLYPASILLRPVAATGNATSTVASPPESIITSLSQQQPQKENIASVNDHLQSVVDRTTNFLLNSSTDLFNAA
LCAAEAQGNAGKTLINAKRARDADDRTHCQRY SAYQTTPIHEQAGQQSPIPVDSRSLHHQVVL PAPPDIMTTQNGYGDLSFSNGGNE SYRNEVNSCLPV SRELPSVQSTESAD
SSSGLSTSFTVTYQMLSLSEVDSFNRS AKCETLGSNDGNLLDVVLQQRDPEFHHLRLKTDGITEQQQQQEAQTLKTYCNCPSRSL SQFSDPLSFLGGRSLTLLSDDLNVAMES
EGVSSDGLLAAGEEVHGEHKQKEVTSSSLCIEQEEDELLTLLAFGR LDSMSFAGLYEEQGGSSDAASMDPDEEFMNSGGGLEAMFANVHNDTATLDSSNTSWEQQQLDSCLLW
NNMPGAVEGRVLLMKHSLRT

In [12]: tree

```
.
├── data
│   ├── Azfi-v1-MYB-sequences.fasta
│   ├── Azfi-v1-MYB-sequences_linear.fasta
│   ├── combi_sequences_linear.fasta
│   ├── MYB33_ARATH.fasta
│   ├── MYB33_ARATH_linear.fasta
│   ├── VII_sequences.fasta
│   ├── VII_sequences_linear.fasta
│   ├── VI_sequences.fasta
│   └── VI_sequences_linear.fasta
├── docs
│   └── step1_differentiate_subfamilies_VI_and_VII.html
├── envs
│   ├── conda-env-jalview.yaml
│   └── conda-env-phylogenetics.yaml
├── figures
│   └── VI-vs-VII_alignment_v1.png
├── step1_differentiate_subfamilies_VI_and_VII.ipynb
└── step2.ipynb
```

4 directories, 15 files

2 align

```
In [13]: conda activate phylogenetics
if [ ! -d ./data/alignments_raw/ ]
then mkdir ./data/alignments_raw
fi
for i in data/*sequences_linear.fasta
do inseq=$(echo $i | cut -d '/' -f 2 | cut -d '.' -f 1)
  if [ ! -f "./data/alignments_raw/$inseq"_aligned-mafft-linsi.fasta ]
  then linsi --thread $(nproc) data/$inseq.fasta > ./data/alignments_raw/"$inseq"_aligned-mafft-linsi.fasta \
    2> ./data/alignments_raw/"$inseq"_aligned-mafft-linsi.log
  fi
done
conda deactivate
```

(phylogenetics) (phylogenetics) (phylogenetics)

```
In [14]: tree data
```

```
data
├── alignments_raw
│   ├── Azfi-v1-MYB-sequences_linear_aligned-mafft-linsi.fasta
│   ├── Azfi-v1-MYB-sequences_linear_aligned-mafft-linsi.log
│   ├── combi_sequences_linear_aligned-mafft-linsi.fasta
│   ├── combi_sequences_linear_aligned-mafft-linsi.log
│   ├── VII_sequences_linear_aligned-mafft-linsi.fasta
│   ├── VII_sequences_linear_aligned-mafft-linsi.log
│   ├── VI_sequences_linear_aligned-mafft-linsi.fasta
│   └── VI_sequences_linear_aligned-mafft-linsi.log
├── Azfi-v1-MYB-sequences.fasta
├── Azfi-v1-MYB-sequences_linear.fasta
├── combi_sequences_linear.fasta
├── MYB33_ARATH.fasta
├── MYB33_ARATH_linear.fasta
├── VII_sequences.fasta
├── VII_sequences_linear.fasta
├── VI_sequences.fasta
└── VI_sequences_linear.fasta
```

1 directory, 17 files

```
In [26]: conda activate jalview
for i in data/alignments_raw/*aligned*.fasta
do prefix=$(echo $i | sed 's/\.fasta//')
jalview -nodisplay \
        -open $prefix.fasta \
        -colour CLUSTAL \
        -png $prefix.png > /dev/null 2> /dev/null
done
conda deactivate
```

(jalview) (jalview)

intermediate conclusions

Already from the alignments, I can see that one Azolla sequence shouldn't be in here so I will remove it. I will remove have removed that and reordered the alignment to place the Azolla sequences in the middle of the two subdomains.

3 Identify domains

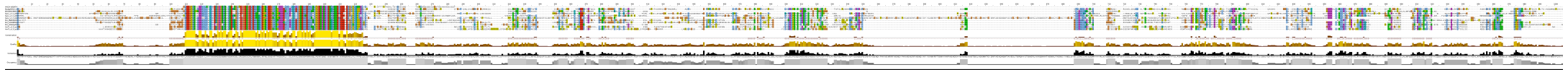
reproducing Jiang & Rao

linsi alignments

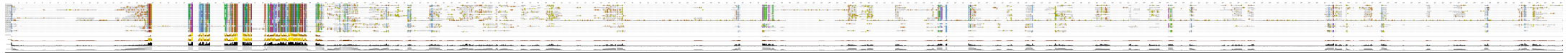
Now let's look at the linsi alignments. If I open them in jalview, I get something like this for VI.



VII



and both combined



There are clearly conserved domains, and especially in VI, there is a super long tail of extra sequence in one of the rows.

And I'm looking for these...

Subfamily VI

- CHBRA15g00250 [CbrDUO1]
- LC221833 [CauDUO1]
- LC221832 [CleDUO1]
- GFZG01001741.1
- Mapoly0019s0071
- XM_024688730.1
- XM_024668389.1
- Sacu_v1.1_s1503.g028048
- Sacu_v1.1_s0147.g023157
- MA_130648g0010
- AmTr_v1.0_scaffold00111.43

22

ARMLASPGDVV
ARLEKQRQRQL
ARLEKQRQRQL
ARMLASPGDVV
LRALQRPKMPS
LRALQRPKTPS
LRALQRPKMGS
LRALQRPKGVD
LRALQRPKGVD
MRALQRPKSQS
ARTLQVAAPLP

Subfamily VII

- DN3051_c0_g1_i1
- HAOX-0009745
- Mapoly1089s0002
- MN199011
- Sacu_v1.1_s0002.g001222
- Sacu_v1.1_s0041.g012546
- Sacu_v1.1_s0272.g027033
- MA_96853g0010
- MA_10208000g0010
- MA_20462g0010
- AmTr_v1.0_scaffold00038.91
- AmTr_v1.0_scaffold00037.85

24

IRQNLPIYPVQLAHF
IRAGLPLHSTDVLVC
MRAGLPVYPPQMNP
MRAGLPVYPPDLQDL
LRAGKPLYPADVKLV
SRAGLPLYPASILLR
ARANLPLYPAEVLAS
QRAGLPLYPPDIQLQ
QRQGLPLYPPDLPLQ
MRAGLSLYPPEVGAQ
QRAGLPLYPPDLHLQ
QRAGLPLYPAEMQRQ

combi alignment from step1

Now let's have a look at the combi alignment, to see if we can actually overlap these regions and use them to differentiate between the two subfamilies.

VII

VI

20	330	340	350	360	370
WNTRIKRRIRAG	---	---	LPLHSTD	---	LVLCPIAT
WNTRIKRRMRAG	---	---	LPVYPPE	---	MQNPA
WNTRIKRRMRAG	---	---	LPVYPPE	---	LQDLCSNLAR
WNTRMKRNL RAG	---	---	KPLYPAD	---	VKL VVKPA
WNTRKRRSRAG	---	---	LPLYPAS	---	ILLR PVAATGNATSTV
WNTRMKRHARAN	---	---	LPLYPAE	---	VVASGAAIAKLEQATW
WNTRIKRRQRAG	---	---	LPLYPPD	---	IQLQN
WNTRIKRRQRQGG	---	---	LPLYPPD	---	LPLOQ
WNTRIKRRMRAG	---	---	LSLYPPE	---	VGAQGQGLDSI
WNTRIKRRQRAG	---	---	LPLYPPD	---	LHLOATGE
WNTRIKRRQRAG	---	---	LPLYPAE	---	MQRQVL SH
WNMRMKKLAKLAR	---	---	LEKQRQRQLL	ISQGSPIAMAAAAMGLPPGA	---
WNMRMKKLAKLAR	---	---	LEKQRQRQLL	ISQGSPIAMAAAAMGLPPGT	---
WNMRMKKLAKLAR	---	---	LEKQRQRQL	---	---
---	---	---	EPMPPPP	L	PQAKACMLAPAVPILVRI
WSTRQKRILRALQ	---	---	RPKMPSG	---	---
WSAMILGF AEAGDAKLALETFRMRMPCAPNR	---	---	---	VTYLAVIKACAALAVVD	---
WSTRQKRLLRALQ	---	---	RPKMGS	---	AAFG
WSTRQKRLLRALQ	---	---	RPKGVDS	---	TSG
WSTRQKRLLRALQ	---	---	RPKGVDS	---	TSG
WSTRQKRIMRALQ	---	---	RPKSQSA	---	---
WSTRQKR LARTLOVA	---	---	APLPPPPPP	---	CKALRPLIGSSEPA

VI

VII

Now let's have a look at the analysis done above but with our two *Azolla* sequences of interest, and one *Arabidopsis* sequence.

	360	370	380	390	400	
VII_HAOX-0009745/1-167	NEIKNFWNTRIKRRIRAG	-----	-----	LPLHSTDLVLC	-----	PIATT
VII_Mp3g05910.1/1-596	NEIKNYWNTRIKRRMRAG	-----	-----	LPVYPPEMQNP	-----	-----
VII_ENA MN199011 MN199011.1/1-435	NEIKNYWNTRIKRRMRAG	-----	-----	LPVYPPDLQDLCSNLARGG	-----	-----
VII_Sacu_v1.1_s0002.g001222/1-653	NEIKNFWNTRMKRNL RAG	-----	-----	KPLYPADVKLV	-----	-----
VII_Sacu_v1.1_s0041.g012546/1-595	NEIKNFWNTRKKRRSRAG	-----	-----	LPLYPASILLR	-----	PVAAT
VII_Sacu_v1.1_s0272.g027033/1-779	NEIKNFWNTRMKRHARAN	-----	-----	LPLYPAEVVAS	-----	GAAIA
VII_MA_96853p0010[582/1-582	NEIKNYWNTRIKRRQRAG	-----	-----	LPLYPPDIQLQ	-----	-----
VII_MA_10208000p0010[602/1-602	NEIKNYWNTRIKRRQRQG	-----	-----	LPLYPPDLPLQ	-----	-----
VII_MA_20462p0010[470/1-470	NEIKNYWNTRIKRRMRAG	-----	-----	LSLYPPEVGAQ	-----	GQGLC
VII_AmTr_v1.0_scaffold00038.91/1-574	NEIKNYWNTRIKRRQRAG	-----	-----	LPLYPPDLHLQ	-----	ATGEG
VII_AmTr_v1.0_scaffold00037.85/1-540	NEIKNYWNTRIKRRQRAG	-----	-----	LPLYPAEMQRQ	-----	VLSHF
sp Q8W1W6 MYB33_ARATH/1-520	NEIKNYWNTRIKRRQRAG	-----	-----	LPLYPPEMHVE	-----	AL---
Azfi_s0004.g008455/1-672	NEIKNFWNTRKKRRLRAG	-----	-----	LSLYPAGVKPN	-----	NGLLS
Azfi_s0021.g015882/1-672	NEIKNFWNTRMKRHIRAR	-----	-----	LPLYPTDVITA	-----	-----
VI_g8575.t1/1-1503	NDVKNFWNMRMKKLAKLAR	-----	-----	LEKQRQRQL	-----	-----
VI_BBG43567/1-154	NDVKNFWNMRMKKLAKLAR	-----	-----	LEKQRQRQL	-----	-----
VI_BBG43566/1-124	NDVKNFWNMRMKKLAKLAR	-----	-----	LEKQRQRQL	-----	-----
VI_Mapoly0019s0071.1/1-402	NDVKNFWS TRQKRILRALQ	-----	-----	RPKMPSGSGTDA	-----	-----
VI_XM_024688730.1/1-386	NDVKNFWS TRQKRILRALQ	-----	-----	RPKTPSGKGDC	-----	-----
VI_XM_024688710/1-588	ERNVVSWSAM LGFAEGDAKLALETFRRMPCAPNRVTYLAVI	-----	-----	-----	-----	-----
VI_XM_024688389.1/1-442	NDVKNFWS TRQKRILRALQ	-----	-----	RPKMGSGAAPG	-----	SCEAS
VI_Sacu_v1.1_s1503.g028048/1-178	NDVKNFWS TRQKRILRALQ	-----	-----	RPKGVDSTS	-----	GVTYG
VI_Sacu_v1.1_s0147.g023157/1-159	NDVKNFWS TRQKRILRALQ	-----	-----	RPKGVDSTS	-----	GVTYG
VI_MA_130648p0010[437/1-437	NDVKNFWS TRQKRIMRALQ	-----	-----	RPKSQS	-----	-----
VI_AmTr_v1.0_scaffold00111.43/1-288	NDVKNFWS TRQKRILARTLQVAAP	-----	-----	LPPPPPPCKAL	-----	-----

When I boldly remove that gap that seems to split up this differentiating domain, it looks like so:

VII_HAOX-0009745/1-167	NEIKNFWNTRIKRRIRAG-LPLHSTDLVLC	---	PIA
VII_Mp3g05910.1/1-596	NEIKNYWNTRIKRRMRAG-LPVYPPPEMQNP	-----	
VII_ENA MN199011 MN199011.1/1-435	NEIKNYWNTRIKRRMRAG-LPVYPPDLQDLCSNLARG	-----	
VII_Sacu_v1.1_s0002.g001222/1-653	NEIKNFWNTRMKRNL RAG-KFLYPADV KLV	-----	
VII_Sacu_v1.1_s0041.g012546/1-595	NEIKNFWNTRKKRRSRAG-LPLYPASILLR	----	PVA
VII_Sacu_v1.1_s0272.g027033/1-779	NEIKNFWNTRMKRHRAN-LPLYPAEVVAS	----	GAA
VII_MA_96853p0010[582/1-582	NEIKNYWNTRIKRRQRAG-LPLYPDILQLQ	-----	
VII_MA_10208000p0010[602/1-602	NEIKNYWNTRIKRRQRQGLPLYPDILQLQ	-----	
VII_MA_20462p0010[470/1-470	NEIKNYWNTRIKRRMRAG-LSLYPPEVGAQ	----	GQG
VII_AmTr_v1.0_scaffold00038.91/1-574	NEIKNYWNTRIKRRQRAG-LPLYPDLHLQ	----	ATG
VII_AmTr_v1.0_scaffold00037.85/1-540	NEIKNYWNTRIKRRQRAG-LPLYP AEMQRQ	----	VLS
sp Q8W1W6 MYB33_ARATH/1-520	NEIKNYWNTRIKRRQRAG-LPLYPPEMHVE	----	AL-
Azfi_s0004.g008455/1-672	NEIKNFWNTRKKRRRLRAG-LSLYPAGVKPN	----	NGL
Azfi_s0021.g015882/1-672	NEIKNFWNTRMKRHIRAR-LPLYP TDMITA	-----	
VI_g8575.t1/1-1503	NDVKNFWMRMKKLAKLARLEKQRQRQL	----	
VI_BBG43567/1-154	NDVKNFWMRMKKLAKLARLEKQRQRQL	-----	
VI_BBG43566/1-124	NDVKNFWMRMKKLAKLARLEKQRQRQL	-----	
VI_Mapoly0019s0071.1/1-402	NDVKNFWS TRQKRILRALQRPKMPSGSTDA	-----	
VI_XM_024688730.1/1-386	NDVKNFWS TRQKRLLRALQRPKTPSGKGDC	-----	
VI_XM_024688710/1-578	ERNVVSWSAMILGF AEAGDMPCAPNRVTYLAVI	----	
VI_XM_024688389.1/1-442	NDVKNFWS TRQKRLLRALQRPKMGSGAAPG	----	SCE
VI_Sacu_v1.1_s1503.g028048/1-178	NDVKNFWS TRQKRLLRALQRPKGVDSTIS	----	GVT
VI_Sacu_v1.1_s0147.g023157/1-159	NDVKNFWS TRQKRLLRALQRPKGVDSTIS	----	GVT
VI_MA_130648p0010[437/1-437	NDVKNFWS TRQKRIMRALQRPKSSQS	----	
VI_AmTr_v1.0_scaffold00111.43/1-284	NDVKNFWS TRQKRLLARTLQLPPPPPC KAL	-----	

Just this alignment indicates that the Azolla sequence of interest belong to the VII subfamily

The XM....710 sequence is still in, I thought I had removed it but somehow I managed to keep it in there... I won't change that now but remove the sequence from the tree instead.

4. Combi alignment with Azolla sequences

as done above

5. phylogeny

Now let's run the phylogenies again on these two new alignments

5.1 trimming

```
In [18]: conda activate phylogenetics
if [ ! -d data/alignments_trimmed ]
then mkdir data/alignments_trimmed
fi

# define appendix only once here:
trimappendix='trim-gt4'

inseq=combi_sequences_linear
for a in "data/alignments_raw/$inseq"_aligned*.fasta
do appendix=$(echo $a | cut -d '/' -f 3- | sed "s/$inseq\_//" | sed "s/\.fasta//")
if [ ! -f data/alignments_trimmed/"$inseq"_"$appendix"_"$trimappendix".fasta ]
then echo "trimming alignment $a"
sed -i 's/ \_ /g' $a
trimal -in $a \
-out data/alignments_trimmed/"$inseq"_"$appendix"_"$trimappendix".fasta \
-gt .4 \
-htmlout data/alignments_trimmed/"$inseq"_"$appendix"_"$trimappendix".html
fi
done
conda deactivate
```

(phylogenetics) (phylogenetics) (phylogenetics) (phylogenetics) (phylogenetics) (phylogenetics) (phylogenetics) tri
mming alignment data/alignments_raw/combi_sequences_linear_aligned-mafft-linsi.fasta
(phylogenetics)

```
In [25]: conda activate jalview
for i in data/alignments_trimmed/*.fasta
do prefix=$(echo $i | sed 's/\.fasta//')
jalview -nodisplay \
-open $prefix.fasta \
-colour CLUSTAL \
-png $prefix.png > /dev/null 2> /dev/null
done
conda deactivate
```

(jalview) (jalview)

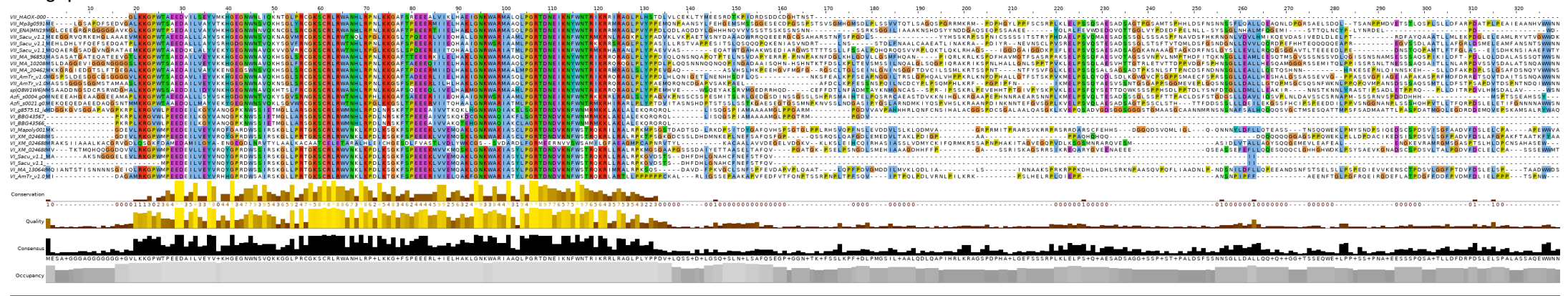
In [22]: tree data/alignments_trimmed

data/alignments_trimmed

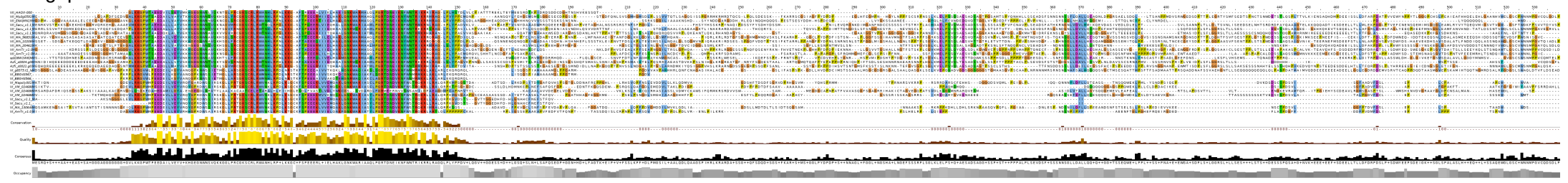
- combi_sequences_linear_aligned-mafft-linsi_trim-gt4.fasta
- combi_sequences_linear_aligned-mafft-linsi_trim-gt4.html
- combi_sequences_linear_aligned-mafft-linsi_trim-gt4.png
- combi_sequences_linear_aligned-mafft-linsi_trim-gt6.fasta
- combi_sequences_linear_aligned-mafft-linsi_trim-gt6.html
- combi_sequences_linear_aligned-mafft-linsi_trim-gt6.png

0 directories, 6 files

For gap threshold 60%



For gap threshold 40%



All alignments contain the differentiating domain as published in J&R. 40% is too gappy for my taste, so I will take the 60% all for tree inference. However, supervisor will like 40% better so I'll run that too.

5.2 tree inference

```
In [27]: inseq=combi_sequences_linear  
echo $inseq
```

```
combi_sequences_linear
```

```
In [28]: ls data/alignments_trimmed/"$inseq"_aligned*gt[46].fasta
```

```
data/alignments_trimmed/combi_sequences_linear_aligned-mafft-linsi_trim-gt4.fasta  
data/alignments_trimmed/combi_sequences_linear_aligned-mafft-linsi_trim-gt6.fasta
```



```

In [29]: conda activate phylogenetics
for a in data/alignments_trimmed/"$inseq"_aligned*gt[46].fasta
do #iqpendix='iqtree-b100'
    iqpendix='iqtree-bb2000-alrt2000'

    echo "making a tree of file $a"
    echo "The first lines of alignment $a look like this"
    head $a

    file_appendix=$(echo $a | cut -d '/' -f 3- | sed "s/$inseq\//" | sed "s/.fasta//")

    if [ ! -d analyses/"$inseq"_trees/"$file_appendix" ]
    then echo "Making a directory $file_appendix to store trees (name based on alignment filename)"
        mkdir -p analyses/"$inseq"_trees/"$file_appendix"
    fi

    iqprefix=analyses/"$inseq"_trees/"$file_appendix"/"$inseq_"$file_appendix_"$iqpendix"
    if [ ! -f "$iqprefix".tree ]
    then nice iqtree -s $a \
        -m MFP \
        -bb 2000 -alrt 2000 \
        -nt AUTO \
        -ntmax $(nproc) \
        -pre "$iqprefix" \
        2> "$iqprefix".stderr \
        > "$iqprefix".stdout
    #cat "$iqprefix".log | mail -s "IQtree_run $a" laura
    fi
done
conda deactivate

```

(phylogenetics) making a tree of file data/alignments_trimmed/combi_sequences_linear_aligned-mafft-linsi_trim-gt4.fasta

The first lines of alignment data/alignments_trimmed/combi_sequences_linear_aligned-mafft-linsi_trim-gt4.fasta look like this

```
>VII_HAOX-0009745
```

```
-----GLKKGPTAEEDVILSEYVMKHGEGN
WNLIQKNTGLPRCGKSCRLRWANHLRPNLKKG-AFSREEEALVIKLHAEIGNKWARMALQ
LPGRDNEIKNFWNTRIKRRIRAG-LPLHSTDVLCPATTTREKLTEYMEESRDTKPID
RDSDDCDGHTNSHVKESSQT-----
```

```
-----
-----
-----
-----
-----
```

Making a directory aligned-mafft-linsi_trim-gt4 to store trees (name based on alignment filename)

making a tree of file data/alignments_trimmed/combi_sequences_linear_aligned-mafft-linsi_trim-gt6.fasta

The first lines of alignment data/alignments_trimmed/combi_sequences_linear_aligned-mafft-linsi_trim-gt6.fasta look like this

```
>VII_HAOX-0009745
```

```
-----GLKKGPTAEEDVILSEYVMKHGEGNWNLIQKNTGLPRCGKS
CRLRWANHLRPNLKKGAFSREEEALVIKLHAEIGNKWARMALQLPGRDNEIKNFWNTRI
KRRIRAGLPLHSTDVLCEKLTMEESRDTKPIDRDSDDCDGHTNST-----
```

```
-----
-----
-----
```

```
>VII_Mp3g05910.1
```

```
ME----LGSAPDFSEDVGALKKGPTSAEDAILVAYVTKHGEGNWSVQKHSGLYRCGKS
CRLRWANHLRPNLKKGAFTPPEERMIIELHAKLGKWARMAAQLPGRDNEIKNYWNTRI
```

Making a directory aligned-mafft-linsi_trim-gt6 to store trees (name based on alignment filename)
(phylogenetics)

```
In [30]: conda activate phylogenetics
for a in data/alignments_trimmed/"$inseq"_aligned*gt[46].fasta
do iqprefix='iqtree-b100'
  #iqprefix='iqtree-bb2000-alc2000'

  echo "making a tree of file $a"
  echo "The first lines of alignment $a look like this"
  head $a

  file_appendix=$(echo $a | cut -d '/' -f 3- | sed "s/$inseq\_//" | sed "s/\.fasta//")

  if [ ! -d analyses/"$inseq"_trees/"$file_appendix" ]
  then echo "Making a directory $file_appendix to store trees (name based on alignment filename)"
    mkdir -p analyses/"$inseq"_trees/"$file_appendix"
  fi

  iqprefix=analyses/"$inseq"_trees/"$file_appendix"/"$inseq"_"$file_appendix"_"$iqprefix"
  if [ ! -f "$iqprefix".tree ]
  then nice iqtree -s $a \
    -m MFP \
    -b 100 \
    -nt AUTO \
    -ntmax $(nproc) \
    -pre "$iqprefix" \
    2> "$iqprefix".stderr \
    > "$iqprefix".stdout
  cat "$iqprefix".log | mail -s "IQtree_run $a" laura
  fi
done
conda deactivate
```

(phylogenetics) making a tree of file data/alignments_trimmed/combi_sequences_linear_aligned-mafft-linsi_trim-gt4.fasta

The first lines of alignment data/alignments_trimmed/combi_sequences_linear_aligned-mafft-linsi_trim-gt4.fasta look like this

>VII_HAOX-0009745

```
-----GLKKGPTAEEDVILSEYVMKHGEGN
WNLIQKNTGLPRCGKSCRLRWANHLRPNLKKG-AFSREEEALVIKLHAEIGNKWARMALQ
LPGRDNEIKNFWNTRIKRRIRAG-LPLHSTDVLCPATTTREKLTEYMEESRDTKPID
RDSDDCDGHTNSHVKESSQT-----
```

```
-----
-----
-----
-----
```

making a tree of file data/alignments_trimmed/combi_sequences_linear_aligned-mafft-linsi_trim-gt6.fasta

The first lines of alignment data/alignments_trimmed/combi_sequences_linear_aligned-mafft-linsi_trim-gt6.fasta look like this

>VII_HAOX-0009745

```
-----GLKKGPTAEEDVILSEYVMKHGEGNWNLIQKNTGLPRCGKS
CRLRWANHLRPNLKKGAFSREEEALVIKLHAEIGNKWARMALQLPGRDNEIKNFWNTRI
KRRIRAGLPLHSTDVLCEKLTMEESRDTKPIDRDSDDCDGHTNST-----
```

```
-----
-----
-----
```

>VII_Mp3g05910.1

```
ME----LGSAPDFSEDVGALKKGPTSAEDAILVAYVTKHGEGNWSVQKHSGLYRCGKS
CRLRWANHLRPNLKKGAFTPEEERMIIELHAKLGNKWARMAAQLPGRDNEIKNYWNTRI
(phylogenetics)
```

5.3 tree results

Uploading the trees to iTOL, see links.

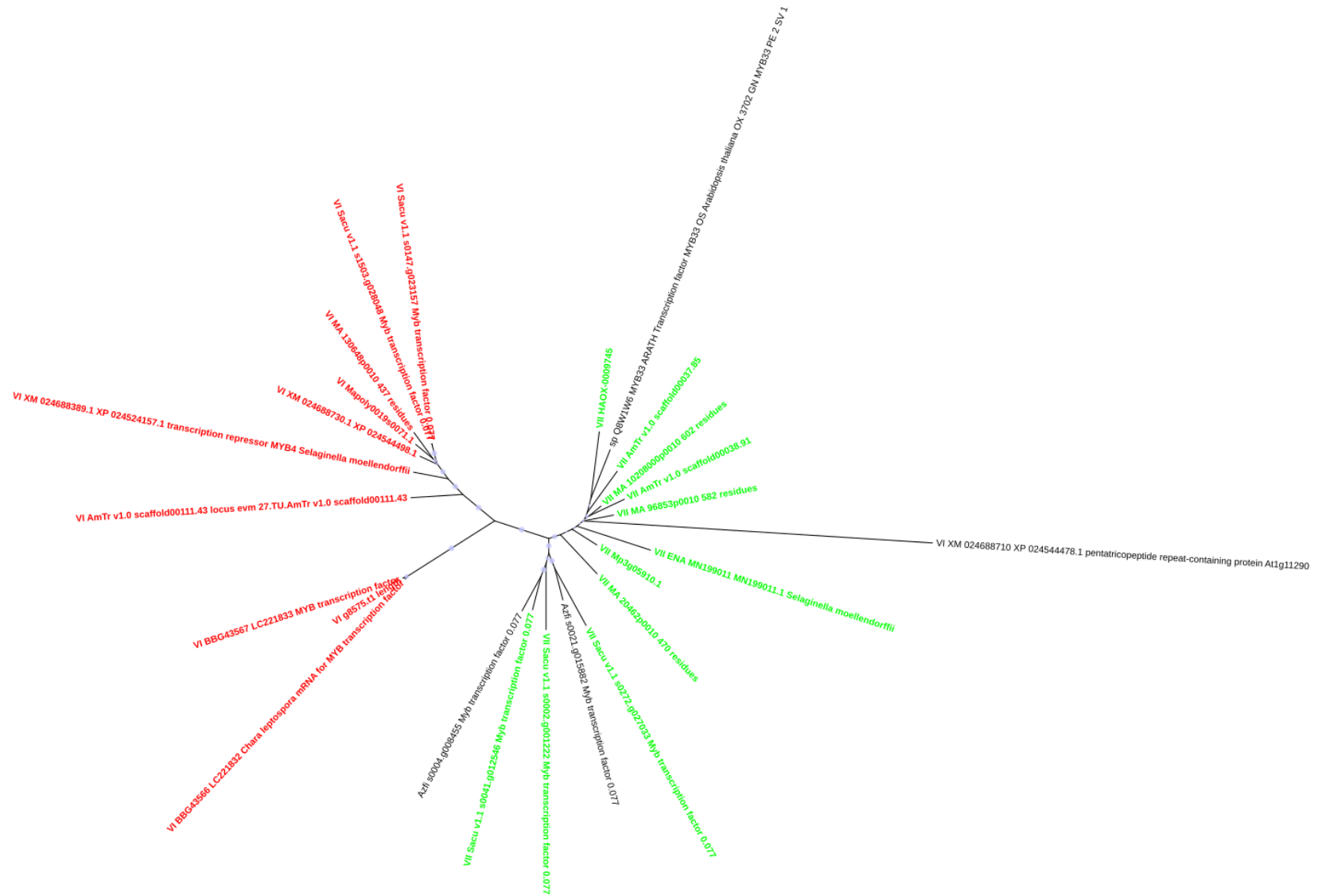
40% gt

Red is VI

Green is VII

gt 40% propper bootstraps:

Tree scale: 10



proper bootstrap (<https://itol.embl.de/tree/1312115964466181596025313>) and ultrafast (<https://itol.embl.de/tree/1312115964273141596034188#>)

60% gt

gt 60% proper bootstraps:

Tree scale: 10

44478.1 petaliocarpale repeat-containing protein K10G1220

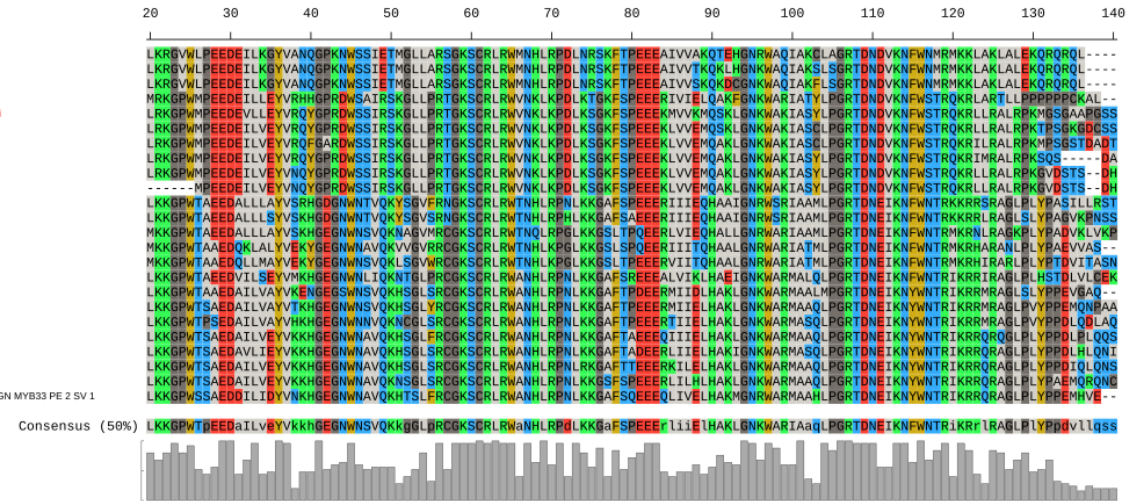
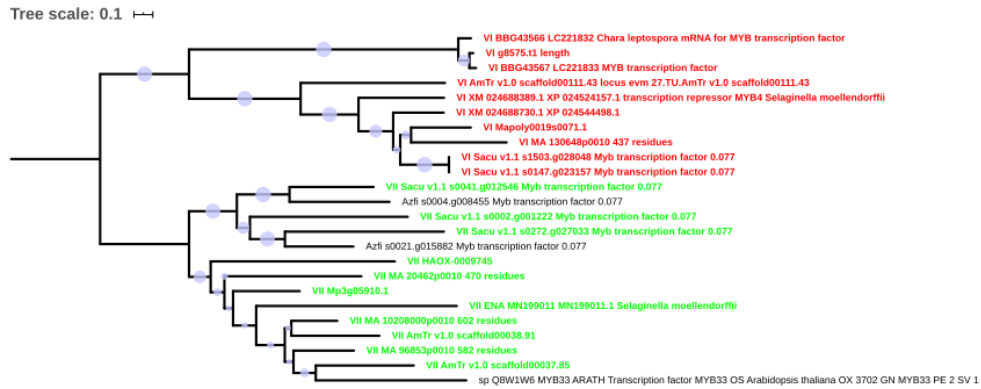
Conclusions

The selected Azolla myb sequences (candidates from RNA seq analysis fit in the VII subfamily of MYBs as identified by R&J. This is supported by (1) inspecting the alignment on the specific area that R&J annotated as characteristic for both sequences. And (2) by the phylogenetic trees. I do wonder if I should include more subfamilies in the phylogeny.

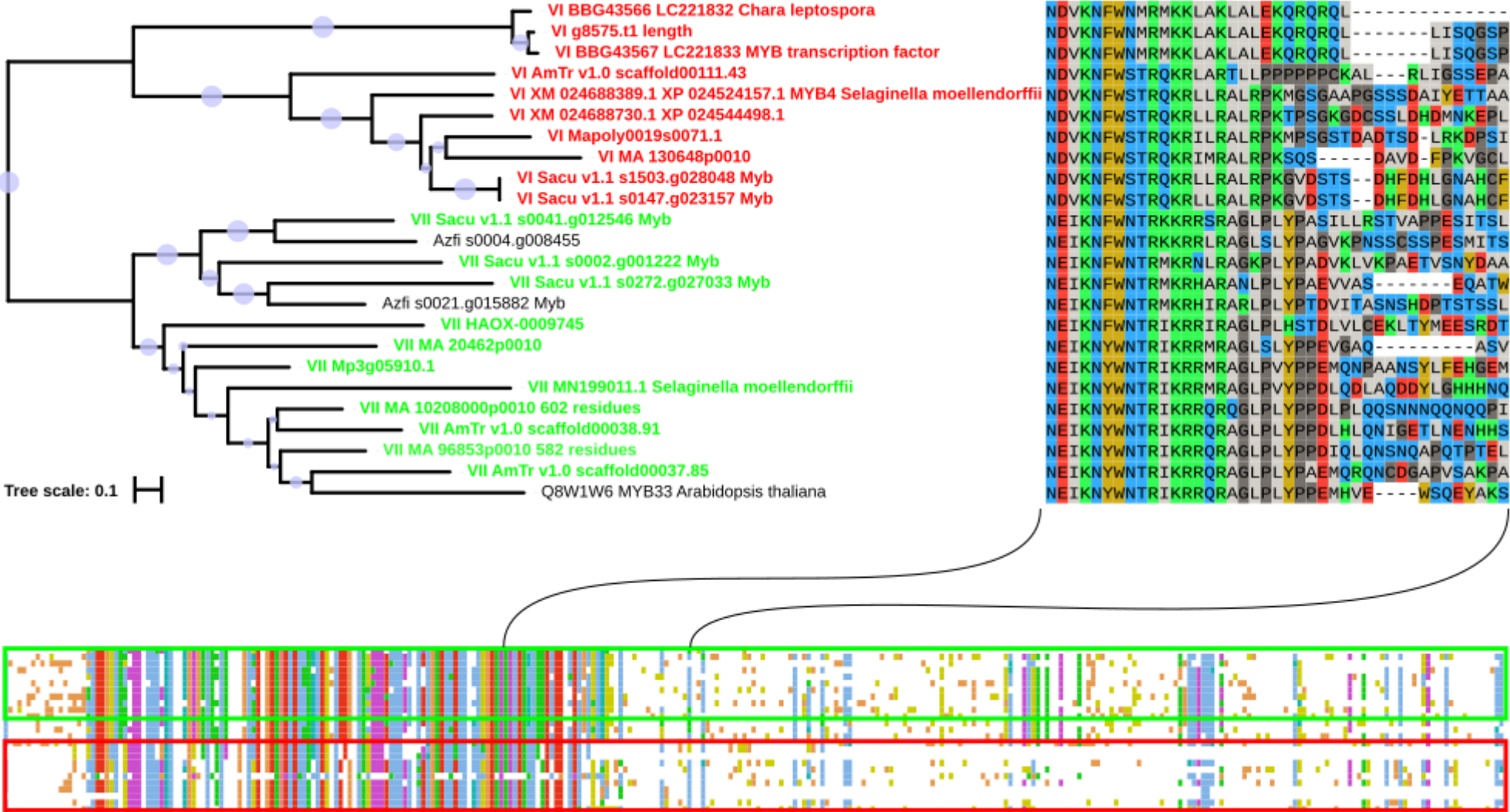
At the very least I need to remove the XM_...710 and the GFZG...1741.1 sequences still, and prepare this into a draft figure.

Another big improvement would be to clean up the sequence names with just an accession nr and add the genus and species names for better interpretation and judgement by the readers.

A figure with the tree and MSA combined could look like this:



Or a bit polished like so:



There's quite some improvements needed still, just thinking out loud here:

1. it's annoying that the two alignments don't have the same colour scheme
2. in the bottom panel, some of the blanks are actually filled with sequence, but they are blank...
3. Better annotation in top right panel of what residues differentiate VI from VII