Anna and I talked and we think that the following list of Genomic resources should be included in the genome browser.

#all in /storage/nr/db/nucleotide/

## all have a built blast database

Genomes

Aegilops longissima acc. 6782:

190708\_Ae\_longissima\_pseudomolecules\_V1.fasta

Aegilops longissima acc. 297:

A297.fas.contigs.fa

Aegilops longissima acc. 1359:

A1359.contigs.fa

Aegilops longissima acc. 1874:

A1874.contigs.fa

Aegilops sharonensis:

181220\_Ae\_sharonensis\_AS1644\_pseudomolecules\_v1.fasta

Aegilops speltoides acc. Tivon:

200420\_Ae\_speltoides\_pseudomolecules\_allChr.fasta

Aegilops tauschii:

Aegilops\_tauschii.Aet\_v4.0.dna.toplevel.fa

Triticum Aestivum (Chinese spring):

160509\_Chinese\_Spring\_v1\_pseudomolecules.fasta

Triticum dicoccoides V2:

zavitan\_nano\_genome\_v2.fa

Triticum durum (Svevo):

160802\_Svevo\_v2\_pseudomolecules.fasta

speltoides annotation:

/storage/raz/ae.annotation/speltoides/ae.speltoides.pgsb.Jul2020.gff3Aegilops

tauschii annotation:

/storage/raz/tauschii/AET\_annotation.gff3

triticum aestivum (Chinese spring):

/storage/raz/iwgsc\_refseq/iwgsc\_refseqv1.0\_2017Mar13.gff3

- Get chromosome size data from the 2bit file

$ twoBitInfo A297.2bit stdout | sort -k2nr > A297\_chromSizes\_sorted.txt

-Build an AGP file from the fasta file, marking all N's as gaps, using the hgFakeAgp command.

$ hgFakeAgp -minContigGap=1 A297.fas.contigs.fa A297.agp

- Check if the new AGP file matches the fasta file:

$ sort -k1,1 -k2n,2n A297.agp > A297\_sorted.agp

$ checkAgpAndFa A297\_sorted.agp A297.2bit

This genome is made of ~8 million contigs and doesn’t have annotation file.

ADD the assembly to the hub

To the genome.txt file in the main folder add:

genome Longissima\_297

trackDb long\_297/trackDb.txt

description Ae\_longissima\_acc\_297

twoBitPath long\_297/A297.2bit

organism Aegilops longissima

scientificName Aegilops longissima

**The genome is actually ~8 million different sequences so no point to browse theone by one.**

**Same for the othe assemblies of longissima**

Upload Aegilops speltoides acc. Tivon:

- Take the fasta sequence and copy it to the gbib data folder:

udiland@icci:~/gbib/gbib\_data/spel$ cp /storage/nr/db/nucleotide/200420\_Ae\_speltoides\_pseudomolecules\_allChr.fasta .

Go to the gbib shell and make a 2bit file:

udiland@icci:~/gbib/gbib\_data/long\_297$ cp /storage/nr/db/nucleotide/A297.fas.contigs.fa .

$ faToTwoBit 200420\_Ae\_speltoides\_pseudomolecules\_allChr.fasta speltoides.2bit

- Get chromosome size data from the 2bit file

$ twoBitInfo speltoides.2bit stdout | sort -k2nr > speltoides.chrom\_sizes\_sorted.txt

-Build an AGP file from the fasta file, marking all N's as gaps, using the hgFakeAgp command.

$ hgFakeAgp -minContigGap=1 200420\_Ae\_speltoides\_pseudomolecules\_allChr.fasta speltoides.agp

- Check if the new AGP file matches the fasta file:

browser@browserbox:/folders/sf\_gbib\_data/spel$ sort -k1,1 -k2n,2n speltoides.agp > speltoides\_sorted.agp

$ checkAgpAndFa speltoides\_sorted.agp speltoides.2bit

speltoides annotation is in:

/storage/raz/ae.annotation/speltoides/ae.speltoides.pgsb.Jul2020.gff3

To upload the annotation track first copy the gff to the folder (using regular shell)

udiland@icci:~/gbib/gbib\_data/spel$ cp /storage/raz/ae.annotation/speltoides/ae.speltoides.pgsb.Jul2020.gff3 .

And than make a genepred file:

$ gff3ToGenePred ae.speltoides.pgsb.Jul2020.gff3 stdout | sort -k2,2 -k4n,4n > spel.genePred

I got two errors one is because there were no gff3 header:

ae.speltoides.pgsb.Jul2020.gff3:1: invalid GFF3 header

The header needs to be: ##gff-version 3

The error can be fixed with:

$ sed -i '1 i\##gff-version 3' ae.speltoides.pgsb.Jul2020.gff3

The second error is because in some rows of mrna description there is no ‘name=’ before the name of the gene. I wrote a python script to fix it.

First I needed to save the outpt of the error so I can take the row number where there is an error

$ gff3ToGenePred ae.speltoides.pgsb.Jul2020.gff3 stdout &> err.txt

(to go to specific line in the file opent it with ‘less’ command and than wright the line number and ‘g’ after it)

Than I ran the fix\_gff.py script.

Now I can make the genePred file:

$ gff3ToGenePred new\_gff2.gff3 stdout | sort -k2,2 -k4n,4n > spel.genePred

Make a bed file:

$ genePredToBed spel.genePred spel.bed

And now make a bigbed file (with the option of name)

$ bedToBigBed -extraIndex=name spel.bed speltoides.chrom\_sizes\_sorted.txt spel.bb

Now enter this genome to the genomes file:

genome Speltoides

trackDb spel/trackDb.txt

groups spel/groups.txt

description 200420\_Ae\_speltoides\_pseudomolecules

twoBitPath spel/speltoides.2bit

organism Aegilpos speltoides acc. Tivon

defaultPos chr1S:1000000-2000000

scientificName Aegilops speltoides

in the same folder as the data add trackDB.txt and groups.txt file:

track genes

type bigBed 12

shortLabel genes aegilpos speltoides

longLabel genes annotations jul2020

bigDataUrl spel.bb

priority 1

visibility packed

groups:

name genes

label genes

priority 1

defaultIsClosed 1

# For the lines of longissima contigs I want to align them to the reference longissima

First I made a bwa index of the reference longissima:

udiland@icci:~/gbib/gbib\_data/long\_297$ bwa index -p long -a bwtsw ../190708\_Ae\_longissima\_pseudomolecules\_V1.fasta

Then I took the 297 line with the contigs and run bwa against the index of the full one

udiland@icci:~/gbib/gbib\_data/long\_297$ cp /storage/nr/db/nucleotide/A297.fas.contigs.fa .

run bwa of longissima 297 contigs against longissimi 6782

udiland@icci:~/gbib/gbib\_data/long\_297$ bwa mem -t 8 long ../../long\_297/A297.fas.contigs.fa -o ../../long\_297/A297\_A6782.sam

convert sam to bam

browser@browserbox:/folders/sf\_gbib\_data/long\_297$ samtools view -S -b -o A297\_A6782.bam A297\_A6782.sam

Sort the bam file

$ samtools sort A297\_A6782.bam A297\_A6782.sorted

Index the bam file

$ samtools index A297\_A6782.sorted.bam

I add this to the 6782 trackDb.txt file :

track Aeg-297

type bam

shortLabel Aeg-297 alignment

longLabel alignment of Aeg-297 against Aeg-6782

bigDataUrl ../long\_297/A297\_A6782.sorted.bam

indelDoubleInsert on

indelQueryInsert on

visibility dense

Now do it to 1359 variant

Copy paste to the folder:

udiland@icci:~/gbib/gbib\_data/long\_1359$ cp /storage/nr/db/nucleotide/A1359.contigs.fa .

run bwa against the 6782 accession

udiland@icci:~/gbib/gbib\_data/long\_1359$ bwa mem -t 8 ../long\_6782/bwa/long A1359.contigs.fa -o A1359\_A6782.sam

From the gbib browser conver to bam and sort and index:

browser@browserbox:/folders/sf\_gbib\_data/long\_1359$ samtools view -S -b -o A1359\_A6782.bam A1359\_A6782.sam

browser@browserbox:/folders/sf\_gbib\_data/long\_1359$ samtools sort A1359\_A6782.bam A1359\_A6782.sorted

browser@browserbox:/folders/sf\_gbib\_data/long\_1359$ samtools index A1359\_A6782.sorted.bam

in A6782 trackDB.txt file:

track Aeg-1359

type bam

shortLabel Aeg-1359 alignment

longLabel alignment of Aeg-1359 against Aeg-6782

bigDataUrl ../long\_1359/A1359\_A6782.sorted.bam

indelDoubleInsert on

indelQueryInsert on

visibility dense

Upload zavitan:

udiland@icci:~/gbib/gbib\_data/zavitan$ cp /storage/nr/db/nucleotide/zavitan\_nano\_genome\_v2.fa .

make 2bit:

browser@browserbox:/folders/sf\_gbib\_data/zavitan$ faToTwoBit zavitan\_nano\_genome\_v2.fa zavitan.2bit

Make chromose sizes file:

browser@browserbox:/folders/sf\_gbib\_data/zavitan$ twoBitInfo zavitan.2bit stdout | sort -k2nr > zavitan.chrom\_sizes\_sorted.txt

mark N as gaps:

browser@browserbox:/folders/sf\_gbib\_data/zavitan$ hgFakeAgp -minContigGap=1 zavitan\_nano\_genome\_v2.fa zavitan.agp

- Check if the new AGP file matches the fasta file:

browser@browserbox:/folders/sf\_gbib\_data/zavitan$ sort -k1,1 -k2n,2n zavitan.agp > zavitan\_sorted.agp

$ checkAgpAndFa zavitan\_sorted.agp zavitan.2bit

This gave an error:

needLargeMem: Out of memory - request size 865950041 bytes, errno: 12

I assume its ok and moved on.

zavitan annotation is in:

/storage/raz/3\_WEW\_v2.0/WEW\_v2.0.gff3

To upload the annotation track first copy the gff to the folder (using regular shell)

udiland@icci:~/gbib/gbib\_data/zavitan$ cp /storage/raz/3\_WEW\_v2.0/WEW\_v2.0.gff3 .

And than make a genepred file:

$ gff3ToGenePred WEW\_v2.0.gff3 stdout | sort -k2,2 -k4n,4n > zavitan.genePred

Got this error:

WEW\_v2.0.gff3:1: invalid GFF3 header

To fix it:

$ sed -i '1 i\##gff-version 3' WEW\_v2.0.gff3

Than I run it again and got this error:

Error: no exon in TRIDC0UG034550.1.mrna1 contains CDS 618752683-618752684

.

.

.

Error: no exon in TRIDC1AG037670.5.mrna1 contains CDS 457031110-457031111

34 errors converting GFF3 file: WEW\_v2.0.gff3

The problem is that in some lines of the gff there are coordination of cds without exon

But the genepred file was created so I continued with the procedure..

$ genePredToBed zavitan.genePred zavitan.bed

$ bedToBigBed -extraIndex=name zavitan.bed zavitan.chrom\_sizes\_sorted.txt zavitan.bb

Now add to genome.txt file:

genome Zavitan

trackDb zavitan/trackDb.txt

groups zavitan/groups.txt

description Triticum dicoccoides V2

twoBitPath zavitan/zavitan.2bit

organism Triticum dicoccoides/zavitan/wild emmer

defaultPos Chr1B:1000000-2000000

scientificName Triticum dicoccoides

Write a trackDb.txt file:

track genes

type bigBed 12

shortLabel genes zavitan

longLabel genes WEW\_V2.0

bigDataUrl zavitan.bb

visibility dense

group genes

Write group.txt file:

name genes

label genes

priority 1

defaultIsClosed 0

Aegilopes tauschii

I downloaded gff from here: <ftp://ftp.ensemblgenomes.org/pub/plants/current/gff3/aegilops_tauschii/>

And the fasta from here:

<ftp://ftp.ensemblgenomes.org/pub/plants/current/fasta/aegilops_tauschii/dna/>

The names of the chromosomes in the fasta are like so:

>1D dna:chromosome chromosome:Aet\_v4.0:1D:1:502330251:1 REF

And the name in the gff are:

1D

So I need to go over the genome and remove everything after the first blank space

I write a script that did it and run it like so:

$ python fix\_fasta\_names.py -in Aegilops\_tauschii.Aet\_v4.0.dna.toplevel.fa -out Aegilops\_tauschii.Aet\_v4.0.dna.toplevel\_fix\_names.fa

Now I can make the needed files:

$ faToTwoBit Aegilops\_tauschii.Aet\_v4.0.dna.toplevel\_fix\_names.fa tauschii.2bit

$ twoBitInfo tauschii.2bit stdout | sort -k2nr > tauschii.chrom\_sizes\_sorted.txt

- Check if the new AGP file matches the fasta file:

$ sort -k1,1 -k2n,2n tauschii.agp > tauschii\_sorted.agp

$ hgFakeAgp -minContigGap=1 Aegilops\_tauschii.Aet\_v4.0.dna.toplevel\_fix\_names.fa tauschii.agp

-Make the annotation track using the gff file

$ gff3ToGenePred Aegilops\_tauschii.Aet\_v4.0.49.gff3 stdout | sort -k2,2 -k4n,4n > tauschii.genePred

$ genePredToBed tauschii.genePred tauschii.bed

$ bedToBigBed -extraIndex=name tauschii.bed tauschii.chrom\_sizes\_sorted.txt tauschii.bb

-Now add to genomes.txt

genome Tauschii

trackDb tauschii/trackDb.txt

groups tauschii/groups.txt

description Aegilops tauschii v4

twoBitPath tauschii/tauschii.2bit

organism Aegilopes tauschii

defaultPos 1D:1000000-2000000

scientificName Aegilopes tauschii

groups tauschii/groups.txt

-Make trackDb.txt file:

track genes

type bigBed 12

shortLabel genes tauschii

longLabel genes tauschii v4.0.49

bigDataUrl tauschii.bb

visibility dense

group genes

-Make groups.txt file:

name genes

label genes

priority 1

defaultIsClosed 0

# split genome to two parts

# because the BLAT tool cannot work with large genome I split genomes to two parts

For sharonensis gnome I used splitfasta to split the genome to individiaual chromosomes and than concat the chromosomes to two files.

Next, I made a .2bit file from each part and made a new entry in the genome.txt file:

$ faToTwoBit 181220\_Ae\_sharonensis\_AS1644\_pseudomolecules\_v1\_chr567un.fasta sharonensis\_chr567un.2bit

genome Sharonensis

trackDb shar/trackDb.txt

groups shar/groups.txt

description 181220\_Ae\_sharonensis\_AS1644\_chr567Un

twoBitPath shar/181220\_Ae\_sharonensis\_AS1644\_pseudomolecules\_v1\_split\_files/sharonensis\_chr567un.2bit

organism Aegilops sharonensis

defaultPos chr5S:1000000-2000000

scientificName Aegilops sharonensis

Than I made the gfServer blat:

# re-atach to screen blat session:

$ screen -r blat

$ ssh browser@localhost -p 1235

$ ~browser/bin/blat/gfServer start 127.0.0.1 42424 -stepSize=5 -log=log/gfServer.log sharonensis\_chr1234.2bit &

Than I add this to the genomes.txt file:

blat 127.0.0.1 42424

isPcr 127.0.0.1 42424

# did the same for the second chromosomes file but there were memory error so I split it to chr56.2bit and chr7Un.2bit and than:

$ ~browser/bin/blat/gfServer start 127.0.0.1 42426 -stepSize=5 -log=log/gfServer\_56.log sharonensis\_chr56.2bit &

$ ~browser/bin/blat/gfServer start 127.0.0.1 42428 -stepSize=5 -log=log/gfServer\_7Un.log sharonensis\_chr7Un.2bit &