Populus Alba-Tremula expression analysis

Ben Sivan

2021-07-07

This is an analysis of RNA sequencing of various tissues from *Populus Alba-Tremula* (PopAT) in order to identify a novel transcription factor that is most representative of a tissue.

The raw data of the tissues is available on NCBI's SRA site.

And the transcriptome assembly for reference from *Populus trichocarpa* (PopTri) available on NCBI's Datasets site.

Retrieving the data is done with the 'wget' bash command.

```
wget -O filename url
```

Extract reads from SRR file using sratoolkit, the output of this command is two fastq files 1 and 2 which are fwd and rev reads respectively.

```
fastq-dump --stdout SRR
```

Process the reads using fastp package (works only on linux).

```
fastp --trim_poly_x --low_complexity_filter --complexity_threshold 50 --thread 16
  -i input_reads1 -I input_reads2 -o output_reads1 -O output_reads1
```

Map reads to reference transcriptome.

```
bwa mem --thread 20 reference.fa read1.fq read2.fq -o alignment.sam
```

Extract only accessions of the mapped reads and then count them.

```
using XAM, BioAlignments

reader = open(SAM.Reader, "data.sam")

io = open("mapped_reads.txt", "w")
for record in reader
        if SAM.ismapped(record)
                print(io, SAM.refname(record), "\n")
        end
end
close(io)
```

```
using FASTX, BioSequences, CSV, DataFrames

Accessions = CSV.read("mapped_reads.txt", DataFrame, header = false)

using Counters

AC = counter(Accessions.Column1)

Accessions_count = DataFrame(Accession = vcat(keys(AC)...), Count = vcat(values(A

CSV.write("Accession_count.csv", Accessions_count)
```

Join all accessions count into one table.

```
Tissues = ["ShootTip", "RootTip", "Bark", "Bud", "Xylem", "Callus", "Leaf"]

AllTissues_dict = Dict()
AllTissues_Accessions = DataFrame(Accession = [])
for tissue in Tissues
        push!(AllTissues_dict, "$tissue" => CSV.read(string(tissue,"_AccessionCou
        rename!(AllTissues_dict[tissue], ["Accession", tissue])
        AllTissues_Accessions = outerjoin(AllTissues_Accessions, AllTissues_dict[
end

for column in names(AllTissues_Accessions)[2:end]
        replace!(AllTissues_Accessions[!,column], missing => 0);
end


# Add the gene description.

description = DataFrame(Accession = [], Description = [])
reader = open(FASTA.Reader, "PopTri_Transcript.fna")

for record in reader
        push!(description, (FASTA.identifier(record), FASTA.description(record)))
end

AllTissues_Accessions = inerjoin(AllTissues_Accessions, description, on = :Access

select!(AllTissues_Accessions, :Accession, :Description, :ShootTip, :RootTip, :Ca

CSV.write("AllTissues_Accessions.csv",AllTissues_Accessions)
```

Normalize by read number per tissue and multiply by 10^8 to resume to convenient numbers.

```
using FASTX

Reads_num = Dict()
for tissue in Tissues
```

```
        cnt = 0
        reader = open(FASTQ.Reader, "$tissue_processed1.fastq")
        for record in reader
                cnt += 1
        end

        push!(Reads_num, tissue => cnt)
end

# Output:
"""
Dict{String,Int64} with 7 entries:
  "Xylem"    => 33379668
  "Bark"     => 45273452
  "Callus"   => 44485862
  "RootTip"  => 55579737
  "ShootTip" => 58336874
  "Bud"      => 58015650
  "Leaf"     => 50610292
"""

for read_num in Reads_num
        AllTissues_Accessions[read_num[1]] = (AllTissues_Accessions[read_num[1]].
end
```

Now is time for the real analysis.

first we will check if the data is representitive as we expect by anelizing the known housekipping genes.

Housekipping genes definition: Housekeeping genes are genes that are required for the maintenance of basal cellular functions that are essential for the existence of a cell, regardless of its specific role in the tissue or organism.

In our context, the genes that their expression is the most consistent, no matter the tissue.

The mathematical way to calculate it is by taking the ratio between standard deviation and the average of each gene per all tissues. Then the smallest number is representative of consistent expression.

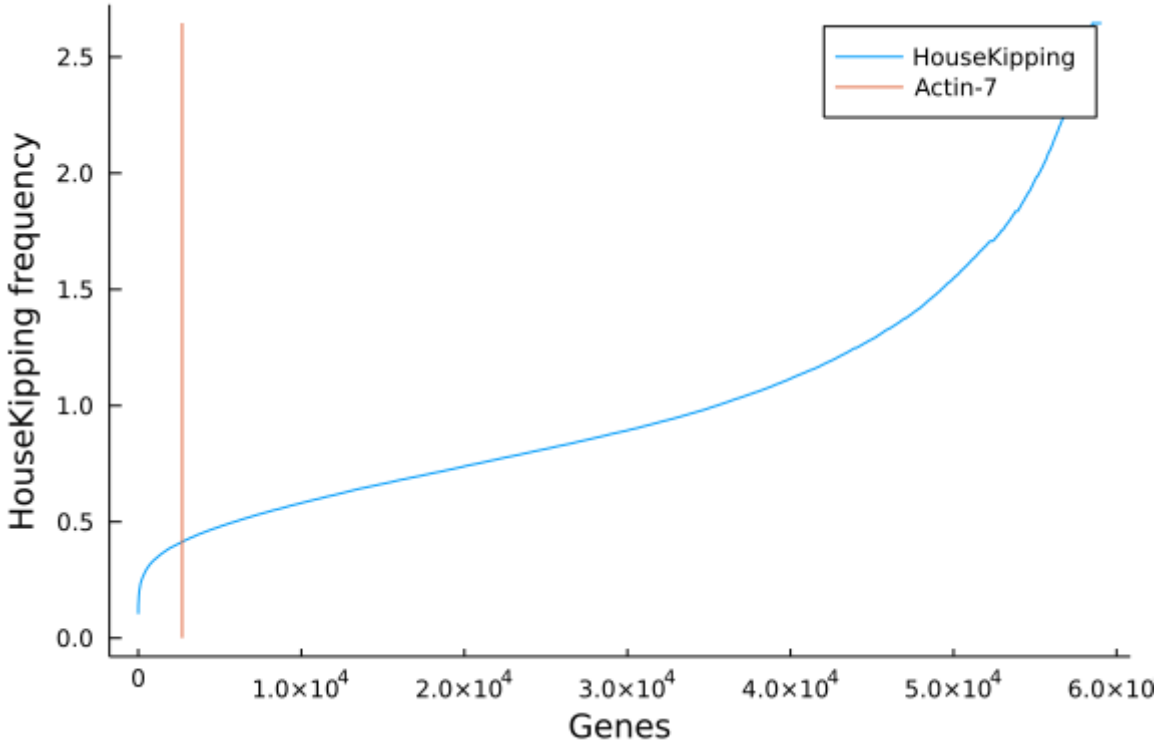
```
using CSV, DataFrames, DataFramesMeta, Statistics, Plots, StatsPlots

AllTissues = CSV.read("AllTissues_Accessions.csv", DataFrame)

select!(AllTissues, :Accession, :Description, :ShootTip, :RootTip, :Callus, :Bud,

AllTissues = @transform(AllTissues, Std = std.(eachrow(AllTissues[:,3:end])), Mea

AllTissues = @transform(AllTissues, SM = :Std./:Mean)

sort!(AllTissues, :SM)

AllTissues[1:5,1:2]
```

5 rows × 2 columns

	Accession	Description
	String	String
1	XR_002978042.1	PREDICTED: Populus trichocarpa uncharacterized LOC112324280 (LOC112324280), ncRNA
2	XM_024597299.1	PREDICTED: Populus trichocarpa E3 ubiquitin ligase PQT3-like (LOC7471635), transcript variant X1, mRNA
3	XM_024604704.1	PREDICTED: Populus trichocarpa 3-hydroxyisobutyryl-CoA hydrolase-like protein 3, mitochondrial (LOC7497640), transcript variant X4, mRNA
4	XM_002312717.3	PREDICTED: Populus trichocarpa uncharacterized LOC7454260 (LOC7454260), transcript variant X1, mRNA
5	XM_002308281.3	PREDICTED: Populus trichocarpa 3-hydroxyisobutyryl-CoA hydrolase-like protein 3, mitochondrial (LOC7497640), transcript variant X1, mRNA

Looks good, now we can plot the distribution of all genes from not tissue specific at all to most specific and we'll put the known housekeeping gene Actin-7 position for reference.

```
function FindGene(name)
        findings = []
        for row in eachrow(AllTissues)
                if occursin(name,row.Description)
                        push!(findings, row.Description)
                end
        end
        return(findings)
end


HouseKipping = @df AllTissues plot(1:nrow(AllTissues),:SM, grid = false, label =

Actin = findfirst(AllTissues.Description .== FindGene("actin-7")[1])

plot!(fill(Actin,nrow(AllTissues)+1), [0:maximum(AllTissues.SM)/nrow(AllTissues):
```



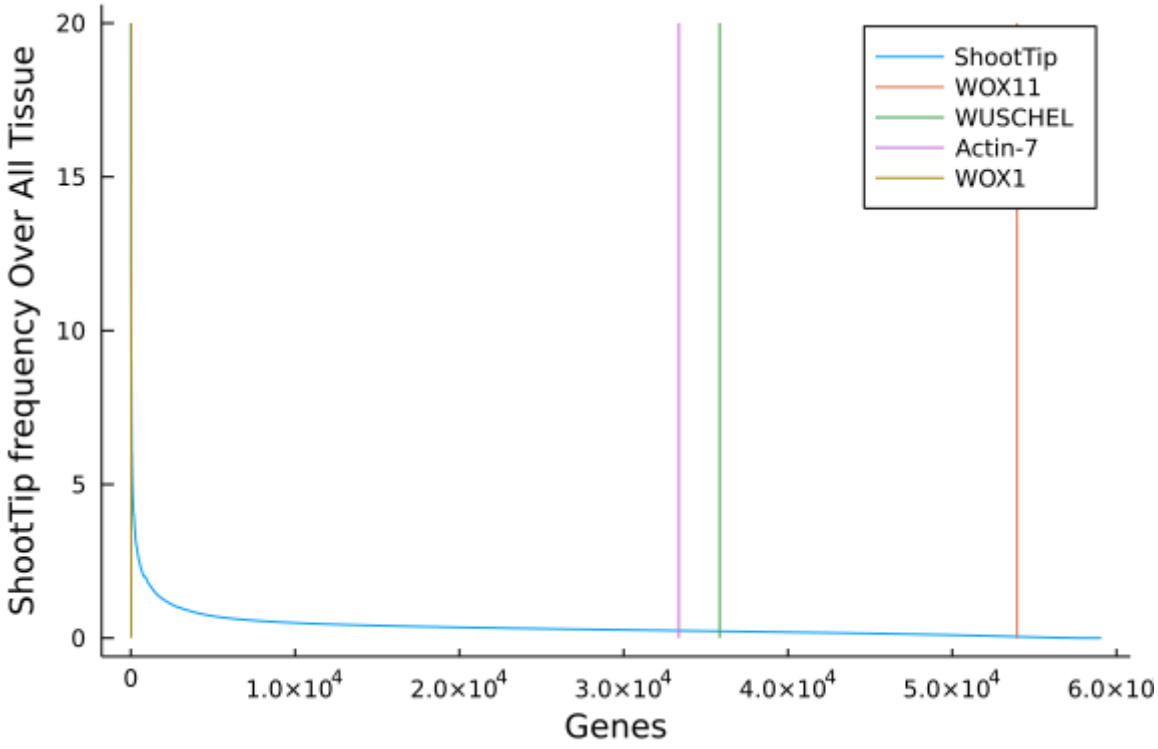
Now that we know the data can show as tissue specific and non tissue specific genes, we can try to identify genes that are representative of our tissue of interest.

In this study, the tissue of interest is the Shoot apical meristem (SAM) and it is located in the shoot tip of the plant.

There are several genes that were identified as SAM specific, such as WUSCHEL (WUS), and WUSCHEL related homeobox (WOX).

In poplar, the gene WOX11 was identified by Bobin et al. as a master regulator for the maintenance of undeveloped cells, like in SAM among others.

```
AllTissues = @transform(AllTissues, ShootOverAll = :ShootTip./(:RootTip.+:Bark.+:

sort!(AllTissues, :ShootOverAll, rev = true)

ShootPlot = @df AllTissues plot(1:nrow(AllTissues),:ShootOverAll, grid = false, l

WOX11 = findfirst(AllTissues.Description .== "PREDICTED: Populus trichocarpa WUSCI

plot!(fill(WOX11,nrow(AllTissues)+1), [0:maximum(AllTissues.ShootOverAll)/nrow(Al

WUSCHEL = findfirst(AllTissues.Description .== "PREDICTED: Populus trichocarpa pr

plot!(fill(WUSCHEL,nrow(AllTissues)+1), [0:maximum(AllTissues.ShootOverAll)/nrow(

Actin = findfirst(AllTissues.Description .== "PREDICTED: Populus trichocarpa acti

plot!(fill(Actin,nrow(AllTissues)+1), [0:maximum(AllTissues.ShootOverAll)/nrow(Al

WOX1 = findfirst(AllTissues.Description .== "PREDICTED: Populus trichocarpa WUSCH

plot!(fill(WOX1,nrow(AllTissues)+1), [0:maximum(AllTissues.ShootOverAll)/nrow(All
```




```
AllTissues[1:5,1:2]
```

5 rows × 2 columns

	Accession	Description
	String	String
1	XM_002300073.1	PREDICTED: Populus trichocarpa agamous-like MADS-box protein AGL103 (LOC7456234), mRNA
2	XM_024593585.1	PREDICTED: Populus trichocarpa protein ACCELERATED CELL DEATH 6-like (LOC112326247), partial mRNA
3	XM_024588316.1	PREDICTED: Populus trichocarpa uncharacterized LOC7495787 (LOC7495787), transcript variant X2, mRNA
4	XR_002976844.1	PREDICTED: Populus trichocarpa probable LRR receptor-like serine/threonine-protein kinase At1g29720 (LOC18103115), transcript variant X3, misc_RNA
5	XM_024581585.1	PREDICTED: Populus trichocarpa probable LRR receptor-like serine/threonine-protein kinase At1g29720 (LOC18103115), transcript variant X2, mRNA

In my analysis, WOX1 was found to be much more shoot-tip specific in comparison to WOX11.

```
FindGene("WUSCHEL-related homeobox 1 ")
```

```
2-element Array{Any,1}:
 "PREDICTED: Populus trichocarpa WUSCHEL-related homeobox 1 (LOC7493492), m
RNA"
 "PREDICTED: Populus trichocarpa WUSCHEL-related homeobox 1 (LOC7462493), m
RNA"
```

Found in positions 29 and 131 out of 59071.

Published from PopAT_Expression.jmd using Weave.jl v0.10.9 on 2021-07-07.