Things you need if you are using windows:

1. Software called 'putty' to connect to Franklin cluster and type in the commands below. X11 forwarding needs to be enabled in order to display the venn diagram. See **Useful commands**.
2. Software called 'winscp' to transfer files to Franklin cluster
3. Software called 'xming' to display x11 program. This is needed to display venn diagram.

There several different analyses you can run:

1. Trim low quality reads and adapter sequences then run quality control (`run_trim_qc.sh`)
2. Align trimmed reads to hg19 or hg38 and create tracks for visualization (`run_align_create_tracks_rna.sh`). Although, this can be run independently of #1, it requires all the files to be in the same path as if #1 has already been run.
3. Count reads in each feature and then performed differential analysis (`run_differential_analysis_rna.sh`). Again, this can be run independently but requires all files to be in their correct location.
4. Instead of running each of the above steps separately, you can run the full RNA-seq analysis (#1-3) using `run_rnaseq_full.sh`
5. After running differential analysis, you can view a venn diagram of differential genes overlap between different groups (`run_overlap.sh`)

Notes:

- User can run create tracks with only 1 replicate but to run full RNA-seq analysis, you need at least two replicates.
- Before running any of these two options, you need to prepare the directory and samples information file called samples.txt by following the steps below.
- You can take a look at `/gpfs0/home/gdlessnicklab/cxt050/Steve/virtual_server/rnaseq-singularity/project_ex` for examples of input files (i.e. samples.txt and fastq directory). `project_ex` folder also contains examples of output files upon completion of running RNAseq full analysis.
- You can also copy the entire `project_ex` folder to your own folder, delete outputs folder and all the *.out files then run this as test files.
- In this document, there are also descriptions of output files.

Preparing directory and samples information

1. You need to have a project directory. This directory can be any name. Please avoid spaces in the name.

2. Inside the project directory, you must have a directory called “fastq”. All fastq files for all replicates/samples should be in this directory. Inside fastq directory, you can have subdirectories which then contain the fastq files. See “project_ex” directory in /gpfs0/home/gdlessnicklab/cxt050/Steve/virtual_server/rnaseq-singularity for example of a project directory.
3. Inside the project directory, you must have a “samples.txt” file which contains your samples information. You can copy the “samples.txt” from “project_ex” to your own project directory, open it with excel and change the text inside to your own samples information. Do **NOT** save the samples.txt as xlsx file then convert it back to txt. Doing so will add extra invisible characters that are not compatible in Unix. When you’re done, save it back as text file with the same name and transfer it to your own project directory. Do **NOT** put empty line(s) in your “samples.txt” in between your samples information.
    a.
        i. **Inside “**samples.txt**”:**
            - **First column: “Groupname”** – type in the group name of the samples. Samples with replicates must have the same “Groupname”. This name must be part of the fastq filename. This column cannot be empty. Do **NOT** use punctuation characters in groupname. Punctuation characters are:

              !"#$%&’()*+,-./:;<=>?@[]^_`{|}~

              Please avoid using a groupname that starts with a number.
                - ❖ **For example**: iEF-714_FLAG_S3_rep1.fastq.gz and iEF-714_FLAG_S3_rep2.fastq.gz which are two biological replicates, both need to have the same name as “Groupname”.
                - ❖ Different groups cannot have “Groupname” that is a part of another group. **For example:** iEF197copy and iEF197 as two different groupnames is NOT ALLOWED!!
            - **Second column: “Controlname”** - type in the name of the control/ background/reference sample. During differential analysis, the corresponding sample will be compared to its “Controlname”. “Controlname” has to be one of “Groupname”. This name must be part of the fastq filename. This column cannot be empty. For control/background sample, please put NA for “Controlname”. Do **NOT** use punctuation characters in controlname. Punctuation characters are:

              !"#$%&’()*+,-./:;<=>?@[]^_`{|}~

              Please avoid using a groupname that starts with a number.
                - ❖ **For example**: “Groupname” “EF714” will have “empty197” as “Controlname” and “empty197” is part of the fastq filename. During differential analysis, fold-change will be calculated with empty197 as reference.

    - Control samples will have NA for “Controlname”. “Controlname” must be one of “Groupname”. **For example**: “Groupname” empty197 will have “NA” as “Controlname”

- **Third column: “Replicatename”** – type in the name of the replicates. This name must be part of fastq filename. This column cannot be empty.
    - **For example:** `A8_20-0647_714R1_S57_L001_R1_001.sub.fastq.gz` will have “R1” as “Replicatename” and `A8_20-0647_714R1_S57_L001_R1_001.sub.fastq.gz` will have “R2” as “Replicatename”
- **Fourth column: “spikename”** – type in NA for all rows.
- **Fifth column: “email”** – type in the email address of the person running the analysis. Slurm job emails will be sent this email address.
- **Sixth column**: “**string_to_identify_sample**” – this string will be used to identify files that below to a particular sample.
    - **For example**: in “project_ex”, there are “`A8_20-0647_714R1_S57_L001_R1_001.sub.fastq.gz`”, “`C8_20-0649_563R1_S59_L001_R2_001.sub.fastq.gz`”,” `H7_20-0646_197R1_S56_L002_R2_001.sub.fastq.gz`” and “`A9_20-0655_197R2_S65_L001_R2_001.sub.fastq.gz`” file names, the “string_to_identify_sample” would be 714R1, 563R1, 197R1 and 197R2 respectively.
    - **Seventh column**: “**string_pair1**” – each fastq file will be a pair of file that only differs in one string. Usually this string is _R1_ for pair1 reads and _R2_ for pair2 reads.
        - For example: in “`project_ex`”, `A9_20-0655_197R2_S65_L001_R1_001.sub.fastq.gz` is pair 1 reads and `A9_20-0655_197R2_S65_L001_R2_001.sub.fastq.gz` is pair2 reads. Therefore the string_pair1 would be “_R1_” and string_pair2 would be “_R2_”
    - **Eight column**: “**string_pair2**” – similar to “**string_pair1**” but for pair 2 reads.

How to

- **Trim adapter sequences and low quality reads then run quality control (QC) without running full analysis.**

After preparing the directory and samples information as detailed above, to trim adapter sequences and low quality reads then run QC without running the full analysis:

1. Go into the project directory.
   - ❖ For example: if your project directory is called “`project_ex`” then type in:
     ```
     cd project_ex
     ```
2. To trim and run QC only, type in:
   ```
   bash /gpfs0/home/gdlessnicklab/cxt050/Steve/virtual_server/rnaseq-
   singularity/scripts/run_trim_qc.sh &> run_trim_qc.out &
   ```

   Check “`run_trim_qc.out`” file for progress and messages.

- **Align reads to reference genome hg19 (or hg38) and create normalized tracks for visualization without running the full analysis.**

  After preparing the directory and samples information as detailed above, to align and create tracks without running the full analysis:

  Note: everything needs to be in the same directories as if run_trim_qc.sh had been run.

1. Go into the project directory.
   - ❖ For example: if your project directory is called “`project_ex`” then type in:
     ```
     cd project_ex
     ```
2. To align to **hg19** and create tracks only, type in:
   ```
   bash /gpfs0/home/gdlessnicklab/cxt050/Steve/virtual_server/rnaseq-
   singularity/scripts/run_align_create_tracks_rna.sh &>
   run_align_create_tracks_rna.out &
   ```

   Check “`run_align_create_tracks_rna.out`” file for progress and messages

3. Or to align to **hg38** and create tracks only, type in:
   ```
   bash /gpfs0/home/gdlessnicklab/cxt050/Steve/virtual_server/rnaseq-
   singularity/scripts/run_align_create_tracks_rna.sh genome=hg38 &>
   run_align_create_tracks_rna.out &
   ```

   Check “`run_align_create_tracks_rna.out`” file for progress and messages

- **Run feature count and differential analysis without running the full analysis.**

  After preparing the directory and samples information as detailed above, to count features and run differential analysis without running the full analysis:

  Note: everything needs to be in the same directories as if run_align_create_tracks_rna.sh had been run.

1. Go into the project directory.

❖ For example: if your project directory is called “`project_ex`” then type in:

```
cd project_ex
```

2. To count features and run differential analysis, type in:

```
bash /gpfs0/home/gdlessnicklab/cxt050/Steve/virtual_server/rnaseq-singularity/scripts/run_differential_analysis_rna.sh &> run_differential_analysis_rna.out &
```

Check “`run_differential_analysis_rna.out`” file for progress and messages.

- ## Run full RNA-seq analysis and obtain lists of differential genes (commonly used)

After preparing the directory and samples information as detailed above, to run the full analysis which includes trim, qc, alignment, tracks creation, feature counting and differential analysis:

1. Go into the project directory.

   ❖ For example: if your project directory is called “`project_ex`” then type in:

```
cd project_ex
```

2. To align to **hg19** and run full RNA-seq analysis (**most commonly used**), type in:

```
bash
/gpfs0/home/gdlessnicklab/cxt050/Steve/virtual_server/rnaseq-singularity/scripts/run_rnaseq_full.sh &> run_rnaseq_full.out &
```

3. Or to align to **hg38** and run full RNA-seq analysis (**most commonly used**), type in:

```
bash
/gpfs0/home/gdlessnicklab/cxt050/Steve/virtual_server/rnaseq-singularity/scripts/run_rnaseq_full.sh genome=hg38 &>
run_rnaseq_full.out &
```

4. Or to specify time limit of 1 day, type in:

```
bash
/gpfs0/home/gdlessnicklab/cxt050/Steve/virtual_server/rnaseq-singularity/scripts/run_rnaseq_full.sh time=1-00:00:00 &>
run_rnaseq_full.out &
```

   Use this command only if you really need the time limit. Any commands described above, can also be used with the time option. This command is usually used for times when the cluster is scheduled for maintenance within a certain time limit. **Note**: your jobs will be killed if they are not finished within the time limit. The format for time is `time=DD-HH:MM:SS`

Check **“`run_rnaseq_full.out`”** file for progress and messages.

Output folders:

In case of errors, you should first read `*.out` in your project folder. For example, if you are running `run_rnaseq_full.sh`, `run_rnaseq_full.out` will contain message about the error and where to look for more information.

After running full RNA-seq analysis, you will have an **outputs folder** with all the results of running the analysis. Inside outputs**, diff_analysis_rslt** folder will have the results of differential analysis. **RNA-seq differential analysis_report.html** which contains detailed report of the entire analysis is a good place to start reading.

bw_files will have the normalized tracks for visualization

fastqc_rslt will have the quality control reports

logs will have logs of programs ran. Details of programs ran will be here

STAR_2pass will have the alignment files produced by running two-pass STAR. The **final bam** files are in "**STAR_2pass/Pass2**". Feature counts used as input for differential analysis are also stored here.

trim will have the trimmed fastq files.

View interactive area-proportional Venn Diagram.

Once you have run differential analysis, you can run Venn diagram app. The app will allow user to change FDR, fold-change, labels, etc. Run the app using the following steps:

1. Go into the project directory.
   - ❖ For example: if your project directory is called "`project_ex`" then type in:
     ```
     cd project_ex
     ```
2. To generate Venn Diagram, type in:
   ```
   bash /gpfs0/home/gdlessnicklab/cxt050/Steve/virtual_server/rnaseq-singularity/scripts/run_overlap.sh
   ```
   Note: There is no '&' at the end of the command!!

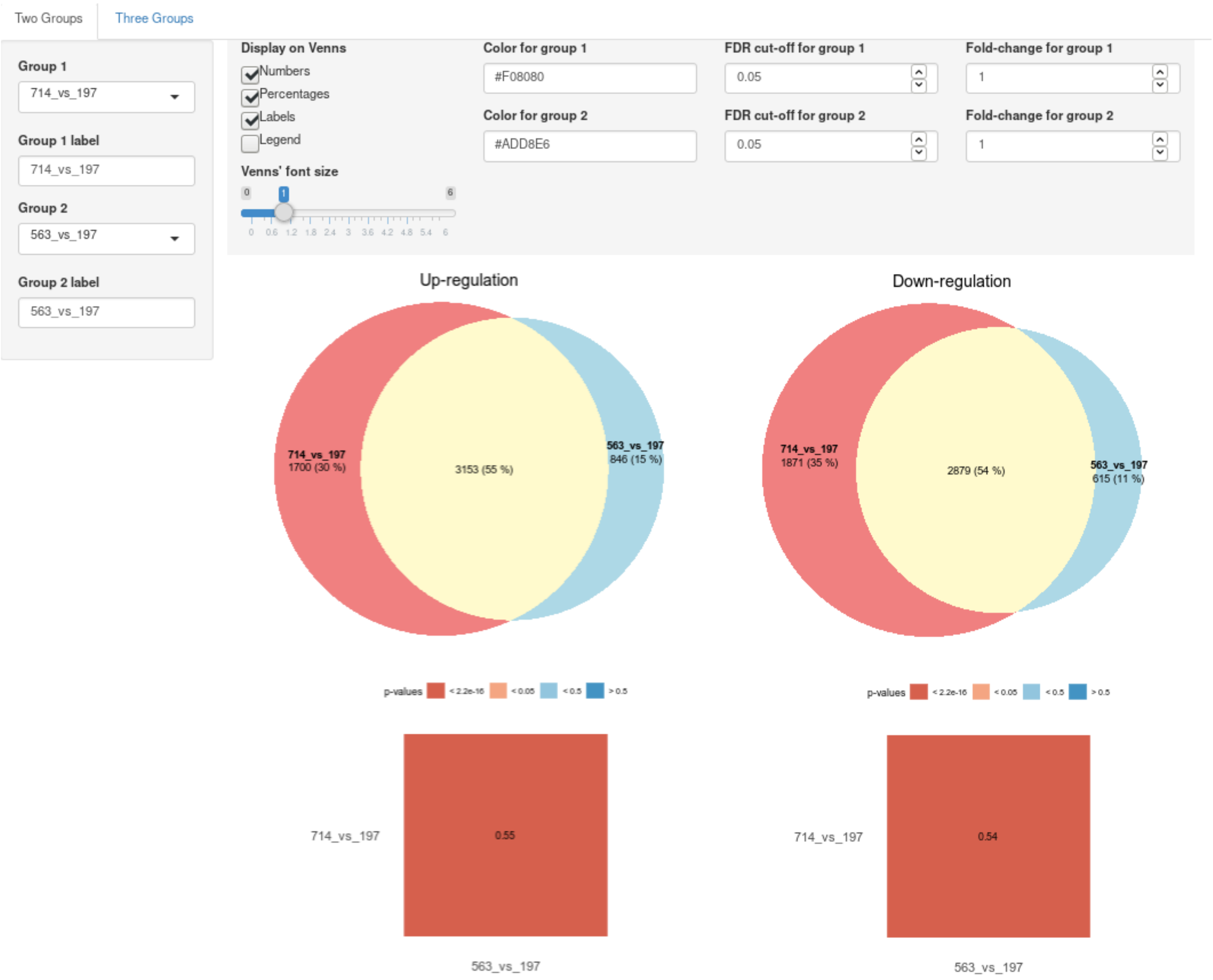
   This will load shiny app, then it will ask for your Franklin cluster's password. After you typed in the password, it will take a while before a firefox with interactive Venn diagram appear. Please wait patiently ☺, it may take a while for firefox to appear.

If it fails with the following error: “connection to … closed by remote host” simply try to re-run the above command.

If a window appears asking you to create new profile, please click on “create new profile”.

Note: you will need to have an x11 display server such as Xming running on your computer (see Useful commands and tips).

Screenshots of the venn diagram app:



Group 1: 714_vs_197

Group 1 label: 714_vs_197

Group 2: 563_vs_197

Group 2 label: 563_vs_197

Display on Venns

☑ Numbers
☑ Percentages
☑ Labels
☐ Legend

Venns' font size: 1

Color for group 1: #F08080

Color for group 2: #ADD8E6

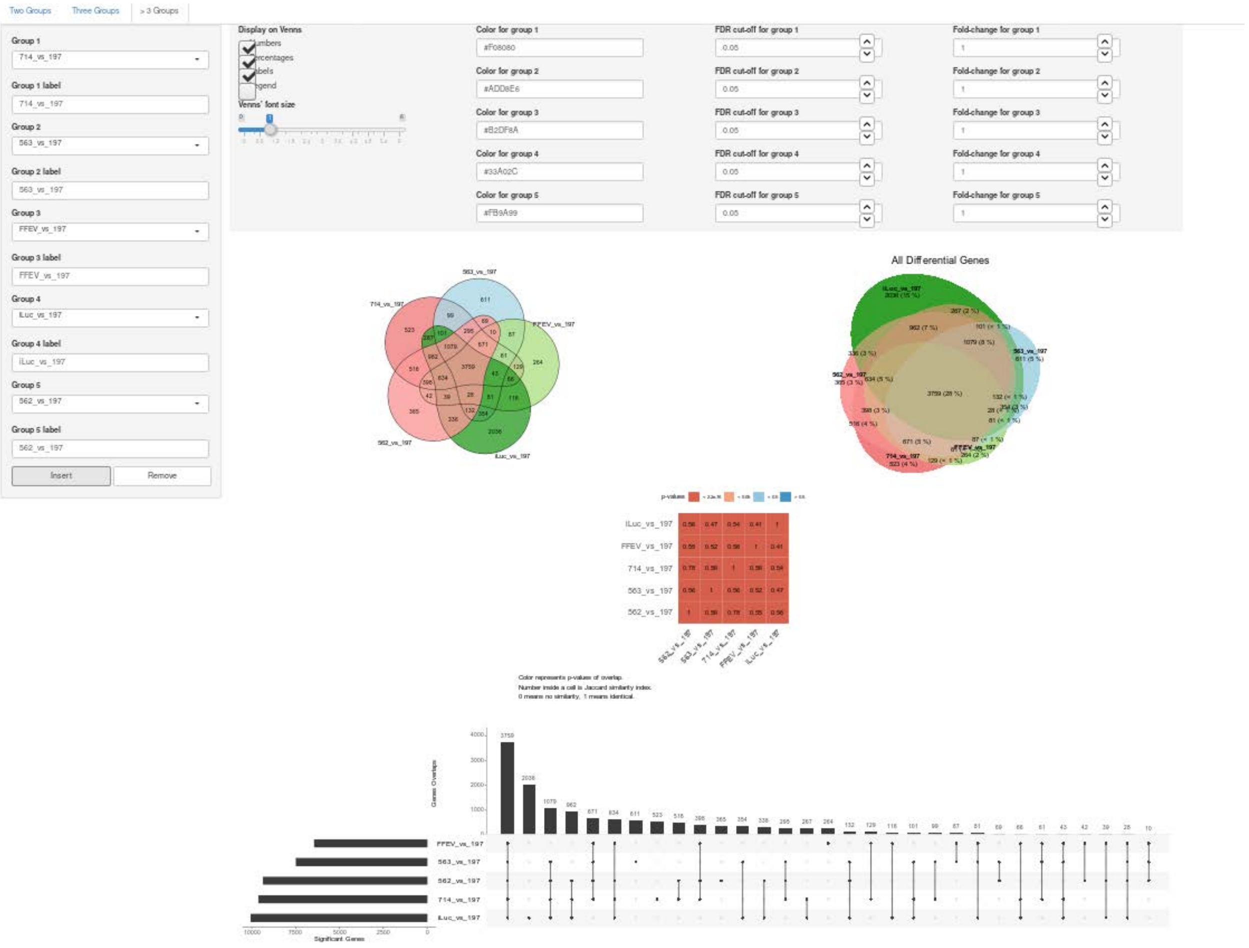
FDR cut-off for group 1: 0.05

FDR cut-off for group 2: 0.05

Fold-change for group 1: 1

Fold-change for group 2: 1

Overlap analysis with more than 3 groups:



Group 1: 714_vs_197

Group 1 label: 714_vs_197

Group 2: 563_vs_197

Group 2 label: 563_vs_197

Group 3: FFEV_vs_197

Group 3 label: FFEV_vs_197

Group 4: iLuc_vs_197

Group 4 label: iLuc_vs_197

Group 5: 562_vs_197

Group 5 label: 562_vs_197

Display on Venns

☑ Numbers

☑ Percentages

☑ Labels

☐ Legend

Venns' font size: 1

Color for group 1: #F08080

Color for group 2: #ADD8E6

Color for group 3: #B2DF8A

Color for group 4: #33A02C

Color for group 5: #FB9A99

FDR cut-off for group 1: 0.05

FDR cut-off for group 2: 0.05

FDR cut-off for group 3: 0.05

FDR cut-off for group 4: 0.05

FDR cut-off for group 5: 0.05

Fold-change for group 1: 1

Fold-change for group 2: 1

Fold-change for group 3: 1

Fold-change for group 4: 1

Fold-change for group 5: 1

	562_vs_197	563_vs_197	714_vs_197	FFEV_vs_197	iLuc_vs_197
iLuc_vs_197	0.56	0.47	0.54	0.41	1
FFEV_vs_197	0.55	0.52	0.56	1	0.41
714_vs_197	0.78	0.56	1	0.56	0.54
563_vs_197	0.56	1	0.56	0.52	0.47
562_vs_197	1	0.56	0.78	0.55	0.56

Color represents p-values of overlap.
Number inside a cell is Jaccard similarity index.
0 means no similarity, 1 means identical.

Useful commands and tips

- To see where you are currently, type:
  pwd
- To change directory type:
  `cd directory_path`

For example:

```
[cxt050@r1pl-hpcf-log01 ~]$ pwd
/gpfs0/home/gdlessnicklab/cxt050
```

Typing 'pwd' let me know that I am currently in "`/gpfs0/home/gdlessnicklab/cxt050`"

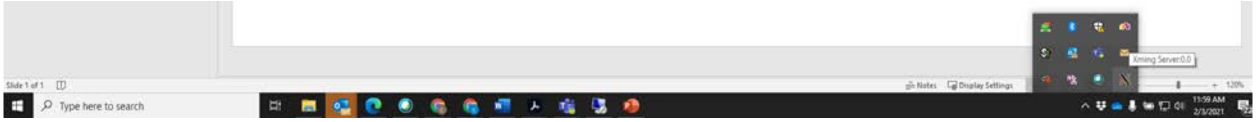
To go into '`/home/gdlessnicklab/share/data/`', type:
"`cd /home/gdlessnicklab/share/data/`"

```
[cxt050@r1pl-hpcf-log01 ~]$ cd /home/gdlessnicklab/share/data/
[cxt050@r1pl-hpcf-log01 data]$ pwd
```

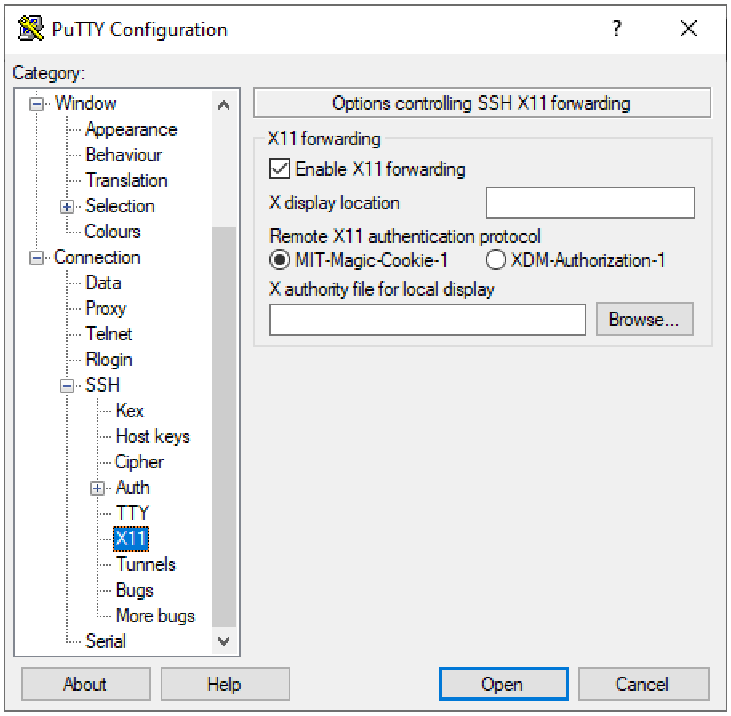
```
/home/gdlessnicklab/share/data
```

To be able to display venn diagram, make sure you have x11 program running:

In the picture below xming (x11 program) is shown running.



If you are using putty, make sure x11 is also checked (enabled):



Options controlling SSH X11 forwarding

X11 forwarding

☑ Enable X11 forwarding

X display location

Remote X11 authentication protocol

● MIT-Magic-Cookie-1 ○ XDM-Authorization-1

X authority file for local display

Have questions? Email cenny.taslim@nationwidechildrens.org