You will need

Paper (or a napkin)

Pen/pencil (preferably two colors)

1. Start with a piece of paper. At the bottom write your hypothesis.
2. Then draw at least 2 graphs. What you will see if your hypothesis is true. What you will see if it is not true.

1. Then write the type of file you will need to draw that graph. (Since this is almost always a csv/tsv file, I like to put the row and column types and names I expect.)
2. Then write program you think can generate that file. (Sometimes this is a custom program you wrote in R or python)

1. Then write the inputs and the outputs to that file.
2. Repeat steps 4-5 until you get back to a fastq file. (If you don’t know what tool to use that this stage that’s ok. Sometimes you try a couple.

Side note:

TO find tools you might want consider

Reading a paper that does something similar to what you want to do

Reading a paper that comparest tools

\*\*\*\*Remember, there is no such thing as the right tool for every job. Only the best tool for your job.

1. You probably need genome files. Where are you going to get them? (don’t get them yet!)
   1. <https://support.illumina.com/sequencing/sequencing_software/igenome.html>
   2. <https://www.ncbi.nlm.nih.gov/genome/>
   3. <https://genome.ucsc.edu/cgi-bin/hgTables>
   4. <https://zoonomiaproject.org/the-mammal-tree-list-view/>

\*\*\*\*\* make sure the fasta file and the gtf/bed match each other!!!

\*\*\*\*\* same number of chromosomes, same chromosome names!!!

* 1. When picking a genome file think about
     1. Would I chase every gene on this list if it matters. If you won’t follow up on a non-coding RNA, don’t pick a list with a non-coding RNA. You have to multiple hypothesis correct for it.

1. Figure out if data already exists, so you don’t have to make it. Or alternatively make the data.
2. Figure out the file structure you are going to use for your project.
   1. Back up raw data and lock the files!!! (chmod)
   2. Back up scripts to github
3. Estimate the size of the data and the size of compute you need.

Figure out which computer you are going to use for your analysis. What does it have installed? Will you have to install things?

1. Go back through the whole plan for quality assurance. Where could you loose data? What effect false positive or false negatives have on your plan? Do you need more controls to make sure the data/tools do not fool you? What quality checks will you do?
2. Share your plan and ask another person to look for issues.

**Now, and only now start analysis.**

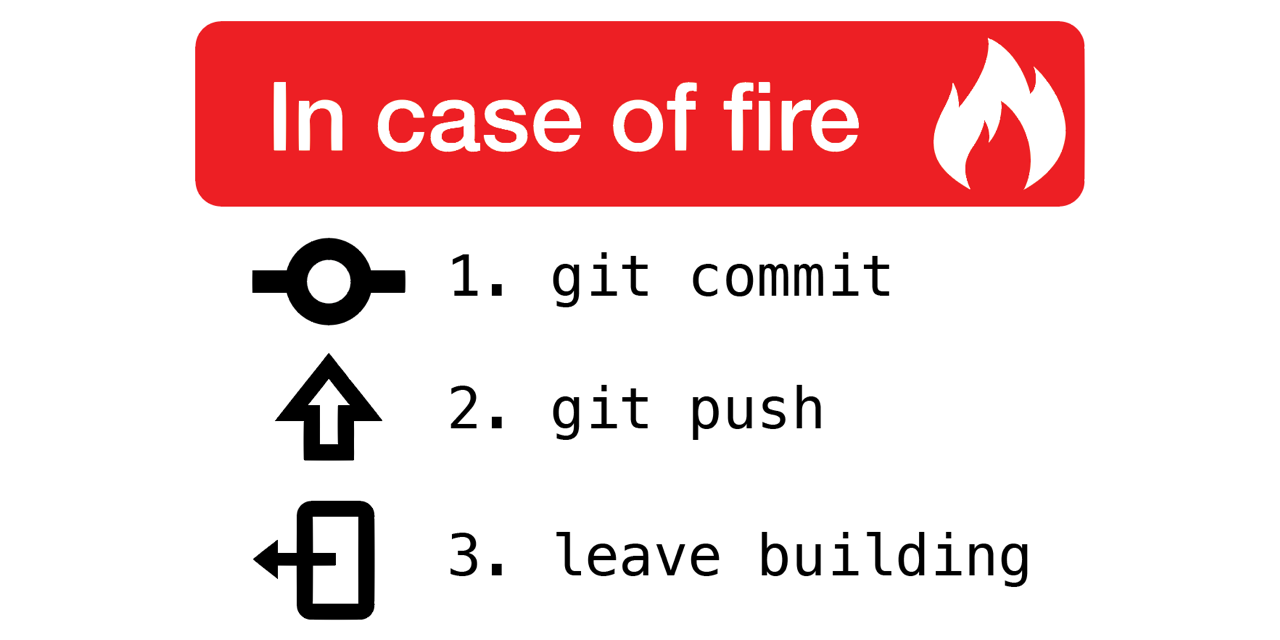
When you download the data, md5 check it!!!!!!!!

When you download the data back it up!!!!

If you don’t have money for backups

<https://www.synapse.org/>

<https://www.ncbi.nlm.nih.gov/geo/>



Virtual environments

### **What is this seq protocol? I've never heard of it!**

[my three-letter seq definition](http://education.knoweng.org/sequenceng/)

### **Where can I learn more about sequencing or bioinformatics?**

[RNA-seq tutorial](https://rnaseq.uoregon.edu/)

[R tutorial (If you are a biologist, the ecology lesson is just as helpful as the genomics lesson. Do both.)](https://datacarpentry.org/lessons/)

[Differential expression of genes](https://github.com/hbctraining/DGE_workshop)

### **Where can I find more about my favorite SNP(s)?**

[Haploreg -connects snps to lots of data](https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php)

### **Where can I find lists of genes that respond to stuff? Or lists of genes that are regulated by stuff?**

[GSEA Gene lists](http://software.broadinstitute.org/gsea/msigdb/genesets.jsp?collection=CGP). List of genes that change in different conditions.

### **Can I find my favorite sequence in some other public data set?**

[**Magic-BLAST**.](https://ncbi.github.io/magicblast/) Blasts a fastq file… but don’t use it for big stuff. Its slow!

[SearchSRA.](https://www.searchsra.org/) You input a fasta file of a seq you care about (like a circlular RNA) and then you input a list of SRRs (samples from GEO) and it can give you back the reads that map against that fasta file from lots of geo samples.

### **Where can I learn more about my favorite gene?**

##### **(Like the expression level in a bunch of different cells types or the SNPs that regulate it, or that perturbations that change its expression, or what regulates my favorite gene)**

[Protien Atlas](https://www.proteinatlas.org/)

[RNA-seq data database](https://www.gtexportal.org/home/)

[fantom CAGE analysis](https://fantom.gsc.riken.jp/cat/v1/#/)

[SSTAR](https://fantom.gsc.riken.jp/5/sstar/Main_Page)

[TFs that regulate this gene (found via chip-seq database)](http://dbtoolkit.cistrome.org/)

[Genecards](https://www.genecards.org/)

[Harmonize](https://maayanlab.cloud/Harmonizome/)

[wiki pathway](https://www.wikipathways.org/index.php/Pathway:WP4117)

[slidebase](http://slidebase.binf.ku.dk/)

[Depmap](https://depmap.org/portal/)

### **Where can I find data on my favorite TF?**

[Chip and ATAC data](http://cistrome.org/db/#/)

[GUI interface overlap chip-seq with public data (chip, sequence, annotation)](https://ccg.epfl.ch/chipseq/chip_extract.php)

### **I did RNA-seq. I got a gene list. What do I do with my gene list? (warning: human or mouse)**

[Enrichr](https://maayanlab.cloud/Enrichr/)

[GSEA tutorial](https://github.com/Dowell-Lab/codeclub/blob/master/gsea/gsea.md)

[DAVID GO terms](https://david.ncifcrf.gov/tools.jsp)

[LISA](http://lisa.cistrome.org/doc)

### **Where can I find more stuff to read? (like you need more stuff to read)**

[Pubmed labs (this is googlefied pubmed searches)](https://www.ncbi.nlm.nih.gov/labs/pubmed/)

[visual pubmed searches](https://pubvenn.appspot.com/)

[word used over time](about:blank)

### **How can I turn one cell type into another?**

[mogrify](http://www.mogrify.net/)

### **Where can I learn more about reproducibility and rigor and responsible conduct of research?**

[Turing way](https://the-turing-way.netlify.app/welcome.html)

[Registered-reports](https://www.cos.io/initiatives/registered-reports)

[Project organization](https://osf.io/)