Meeting with Paolo

-So Paolo seems to think a single transcript/exon analysis is NOT the way to go…tend to miss information (ie. Nodes in our overlap graph) this way
-Also thinks that Olga's software won't do the job….both of them thinking the usefulness of this type (ie. Exon by exon) of software is extremely limited
-better approach: do assembly of ENTIRE transcriptome
-Paolo developed new software for this ^, which he seems to think is better than anything else out there
-Not just for AS type analysis, but for transcriptome assembly in general
-Big idea here: his software looks for overlapping K-mers w/in BOTH reads of paired-end set
-Idea is that we minimize graph complexity this way….ie. the number of uncertainties, or bubbles in our graph
-Most software will initially form a hella complex graph, and then go to great lengths to detangle it
-Graph is constructed of k-mers (i.e.. 16 bp) of size smaller than actual read length
-After graph construction, you have a very messy graph, which we clean up by mapping on the actual reads, which are hopefully long enough to resolve bubbles
-Two major sources of graph complexity:
A) Repetitive/complex regions longer than K-mers
B) Repetitive/complex regions longer than actual reads
-So even after graph cleaning, things are still gonna be very messy
-Here's where PacBio reads come in!
-De Bruijn graph vs overlap graph based assembly
-Here we're doing overlap graph based

————————————————————————————————
FancyPantz Transcriptome Assembler

-Need to run from Ubuntu instance, on AWS cloud
-Are we doing *de novo* assembly here? Are all transcriptome assemblies *de novo*?
-Trinity definitely is….
-Need dir called /reads
-Put everything else in a dir
-.so file -> shared object file….actual code is in c++, but we have some sort of a python wrapper
-Whats actually in Spyros-3cells.py?
-We're not actually doing anything in here?
-Lowercase bases -> lower coverage
-what commands is he actually running??
-kmer length -> this is a param, can be changed….how much overlap do you have between
-Paolo away 20-27th
-but here this week!
-hashFraction variable -> the higher the better, but memory requirement goes up
-docker vs virtualBox -> ubuntu environments for Mac
-bloom filter?
-python memory mapping?
-are you supposed to run all of these options one after another?? They look like steps…
-single cell RNAseq reads/cell vs bulk seq??
-look into: chanzuckerberg-docker (GitHub)
-gpu cluster?

————————————————————————————————

STEPS
> import cziRNA1.so
will make reads binary file?
this could be really huge….be careful
>spyros-3cells.py 1

1 is an option flag
not sure about the difference between options here

```
>spyros-3cells.py 2
>spyros-3cells.py 3 …etc.
```

Questions (for Olga)

-What is a standard object (.so) file?

-Is Spyros-3cells.py a wrapper function for his code? If so, then where is his code (c++) actually located?

-Are the option flags intended to be run one after another? Is it a series of steps?