FGC/NGSC : Answers to Frequently Asked Questions

Jonathan Schug - Technical Director
Klaus Kaestner - Director

... Report problems, errors, or questions to us!

June 14, 2014
Table of Contents

For People in a Big Hurry
  Quick Start
  How to Find Us

Just Enough Detail
  Sequencing at UPenn
  Getting Started
  Costs
  Bringing Samples
  Preparing Samples
  Documenting Samples
  How’s It Going

Many Details
  The Website
  Microarrays
  *-Seq
  Sample Quality Checks
This document tries to cover all of the basic questions that our clients need to know before beginning a high-throughput sequencing experiment or while handling the data produced by the FGC or NGSC cores.

Sequencers
We have 4 hiSeq2000 sequencers, a hiSeq2500, and a miSeq.

Scheduling Sequencing
The hiSeq2000 sequencers sequence eight lanes and we put together sets of eight lanes to fill the sequencers. You do not have to have eight lanes worth to sequence with us.

Bringing In Samples
1. Make sure you have a confirmed investigation set up first.
2. Check calendar to see when we can accept samples.
3. Bring in all of the sample you have. We will return what we do not use.
4. Do not predilute.
5. Read further in this document for more info.

When Will My Sample Be Sequenced?
Check the sample queue to see how your samples fit into all the samples we have.
Check the Runs Calendar to see if we have scheduled it for sequencing and/or when the first open slot is.

Common URLs
Here is how to get in touch with us.

- NGSC Home Page: http://ngsc.med.upenn.edu
- FGC Home Page: http://fgc.genomics.upenn.edu
- Billing Downloads: http://ngsc.med.upenn.edu/Billing/PI
- Data Downloads: Download Data button
- Email: ngsc@mail.med.upenn.edu
For Locals

- We are located on the 12th floor of the Smilow Center.
- See the map below.
- Check the Lab Calendar at the website for samples drop off times

For Distant Collaborators

- Ship overnight in plenty of dry ice.
- Do not ship on Friday!
- The shipping address is:

Room 12-160
Smilow Center for Translational Research
3400 Civic Center Blvd Bldg 421
Philadelphia, PA 19104-5156
Table of Contents

For People in a Big Hurry
  Quick Start
  How to Find Us

Just Enough Detail
  Sequencing at UPenn
  Getting Started
  Costs
  Bringing Samples
  Preparing Samples
  Documenting Samples
  How’s It Going

Many Details
  The Website
  Microarrays
  *-Seq
  Sample Quality Checks
Which Core Do I Want?

There are a few core facilities at Penn that do DNA sequencing. The NGSC has 3 Illumina hiSeq2000, hiSeq2500, and a miSeq. Here is what these machines are good for.

**hiSeq2000**

The hiSeq2000 is good for the techniques **RNA-Seq, ChIP-Seq, miR-Seq, HITS-CLIP, exome capture, BIS-Seq, and whole genome** sequencing in mammals (and their many variations).

There are two aspects of ultra-high throughput sequencing that are important **counts and coverage**. Counts are important for RNA-Seq, ChIP-Seq, miR-Seq, and HITS-CLIP. Coverage is important for exome capture, BIS-Seq, and whole genome sequencing.

The hiSeq2000 generates sequence for about **200 million** fragments per lane. For each fragment the hiSeq can produce single or paired-end 50bp or 100bp sequences. Using 100bp pair-end sequencing, you can get up to 40Gb per lane.

In many cases a single lane can generate more counts or coverage than a sample needs. In that case, it is important to use **multiplexed adapters** so we can sequence multiple samples per lane. Multiplexing samples in a pool is generally a good idea is they allow samples to be test sequenced for quality, then sequenced deeper as needed. It also eliminates any potential for lane-to-lane variations.

**miSeq**

The miSeq uses the same libraries as the hiSeq2000. It generates only about 15 million fragments per lane, but runs very quickly, and can generate reads as long.

**hiSeq2500**

The hiSeq2500 is similar to the hiSeq2000 but has a rapid mode as well as the standard high-output mode. We only run the hiSeq2500s in rapid mode.

The rapid mode uses a different flow cell which has only 2 lanes. If the same sample is placed in both lanes, then clustering is done on the sequencer and the whole process is very quick. Two different samples can be placed in each lane if need be.

Our hiSeq2500s were acquired in 2011 (or so) so they produce somewhat fewer reads per lane than the modern hiSeq2000s.

In rapid mode, the hiSeq2500 can generate 150PE reads which the high-output mode can not do.

**hiSeq Statistics**

<table>
<thead>
<tr>
<th>Technique</th>
<th>Typical Volume</th>
<th>Samples per Lane</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-Seq</td>
<td>30 to 200 million reads</td>
<td>1 to 6</td>
</tr>
<tr>
<td>ChIP-Seq</td>
<td>30 to 100 million reads</td>
<td>2 to 6</td>
</tr>
<tr>
<td>miR-Seq</td>
<td>10 million reads</td>
<td>20</td>
</tr>
<tr>
<td>HITS-CLIP</td>
<td>30 million reads</td>
<td>6</td>
</tr>
<tr>
<td>Exome capture</td>
<td>20-30x coverage</td>
<td>5 to 20</td>
</tr>
<tr>
<td>BIS-Seq</td>
<td>20-30x coverage</td>
<td>1/3</td>
</tr>
<tr>
<td>Genome Sequencing</td>
<td>20-30x coverage</td>
<td>1/3</td>
</tr>
</tbody>
</table>
How Do I Get Started?

The basic things you need to do

1. Read the other FAQs about experiment design and library prep
2. If you still have questions, check the Consultation Calendar and propose a time to meet to discuss your questions.
3. Have your PI create an account (this only needs to be done once.)
4. Create an account for yourself (this only needs to be done once.)
5. Submit an experiment request (do this once for each investigation or for changes to an investigation.)
6. Every weekday morning we review experiment requests, contact you to resolve any questions, then load the design into our database.
7. We contact your BA to make sure everyone is ready to accept the estimated charges when they occur.
8. We will notify you that we can accept samples.
9. Bring samples by during the sample drop-off times posted on the Lab Calendar on the front page of the website.

Fine Print

- If necessary, meet with the Technical Director (Jonathan Schug) to help design the experiment.
  a. Use the BioInfo Calendar tab to identify a time to meet.
  b. Send him an email (jschug@upenn.edu) to propose the time.
How Do I Get Started?

The basic things you need to do

1. Read the other FAQs about experiment design and library prep
2. If you still have questions, check the Consultation Calendar and propose a time to meet to discuss your questions.
3. Have your PI create an account (this only needs to be done once.)
4. Create an account for yourself (this only needs to be done once.)
5. Submit an experiment request (do this once for each investigation or for changes to an investigation.)
6. Every weekday morning we review experiment requests, contact you to resolve any questions, then load the design into our database.
7. We contact your BA to make sure everyone is ready to accept the estimated charges when they occur.
8. We will notify you that we can accept samples.
9. Bring samples by during the sample drop-off times posted on the Lab Calendar on the front page of the website.

Fine Print

- Create accounts at the NGSC or FGC website using the Create New Account link.
  a. The PI of the lab should create the first account.
  b. The investigator(s) who will actually do the experiment should then create their account(s).
  c. Creating a PI’s account works best on Safari on a mac - we are fixing this!
  d. Investigators should make sure they pick their PI from the list when creating their account.
  e. We will active the accounts within a day or so.
How Do I Get Started?

The basic things you need to do

1. Read the other FAQs about experiment design and library prep
2. If you still have questions, check the Consultation Calendar and propose a time to meet to discuss your questions.
3. Have your PI create an account (this only needs to be done once.)
4. Create an account for yourself (this only needs to be done once.)
5. Submit an experiment request (do this once for each investigation or for changes to an investigation.)
6. Every weekday morning we review experiment requests, contact you to resolve any questions, then load the design into our database.
7. We contact your BA to make sure everyone is ready to accept the estimated charges when they occur.
8. We will notify you that we can accept samples.
9. Bring samples by during the sample drop-off times posted on the Lab Calendar on the front page of the website.

Fine Print

▶ Create an experiment using the Create New Experiment button.
  a. This is a basic description of the experiment you want to do.
  b. The template indicates the info we are looking for. If you do not know an answer, say so or go back to Step 2 above.
  c. You need to include your billing information.
  d. We will generate an estimate of the total charges and confirm with your BA that the account can support the charges.
How Do I Get Started?

The basic things you need to do

1. Read the other FAQs about experiment design and library prep
2. If you still have questions, check the Consultation Calendar and propose a time to meet to discuss your questions.
3. Have your PI create an account (this only needs to be done once.)
4. Create an account for yourself (this only needs to be done once.)
5. Submit an experiment request (do this once for each investigation or for changes to an investigation.)
6. Every weekday morning we review experiment requests, contact you to resolve any questions, then load the design into our database.
7. We contact your BA to make sure everyone is ready to accept the estimated charges when they occur.
8. We will notify you that we can accept samples.
9. Bring samples by during the sample drop-off times posted on the Lab Calendar on the front page of the website.

Fine Print

- We will formalize the experiment description and load it into the database.
  a. This process takes a few days.
  b. We may send emails to clarify details of the experiment.
  c. We will also contact your BA to confirm that the budget code is correct and is can support the estimated charges.
How Do I Get Started?

The basic things you need to do

1. Read the other FAQs about experiment design and library prep
2. If you still have questions, check the Consultation Calendar and propose a time to meet to discuss your questions.
3. Have your PI create an account (this only needs to be done once.)
4. Create an account for yourself (this only needs to be done once.)
5. Submit an experiment request (do this once for each investigation or for changes to an investigation.)
6. Every weekday morning we review experiment requests, contact you to resolve any questions, then load the design into our database.
7. We contact your BA to make sure everyone is ready to accept the estimated charges when they occur.
8. We will notify you that we can accept samples.
9. Bring samples by during the sample drop-off times posted on the Lab Calendar on the front page of the website.

Fine Print

▶ Once we have loaded the experiment, bring your samples to the core.
  a Please do not bring the samples until the experiment has been loaded.
  b Be prepared to provide the following for each sample:
    i a source ID
    ii a sample name
    iii the barcode and pool name for multiplexed samples
    iv sequencing dimensions
How Do I Get Started?

The basic things you need to do

1. Read the other FAQs about experiment design and library prep
2. If you still have questions, check the Consultation Calendar and propose a time to meet to discuss your questions.
3. Have your PI create an account (this only needs to be done once.)
4. Create an account for yourself (this only needs to be done once.)
5. Submit an experiment request (do this once for each investigation or for changes to an investigation.)
6. Every weekday morning we review experiment requests, contact you to resolve any questions, then load the design into our database.
7. We contact your BA to make sure everyone is ready to accept the estimated charges when they occur.
8. We will notify you that we can accept samples.
9. Bring samples by during the sample drop-off times posted on the Lab Calendar on the front page of the website.

Fine Print

- We will check the quality of the samples and report any problems. We will not sequence a sample until we have your acknowledgement.
  a. We will schedule the libraries for sequencing and report our progress.
  b. FASTQ files can be downloaded via the website.
  c. Data analysis will follow sequencing as soon as possible.
How Do I Get Started?

The basic things you need to do

1. Read the other FAQs about experiment design and library prep.
2. If you still have questions, check the Consultation Calendar and propose a time to meet to discuss your questions.
3. Have your PI create an account (this only needs to be done once.)
4. Create an account for yourself (this only needs to be done once.)
5. Submit an experiment request (do this once for each investigation or for changes to an investigation.)
6. Every weekday morning we review experiment requests, contact you to resolve any questions, then load the design into our database.
7. We contact your BA to make sure everyone is ready to accept the estimated charges when they occur.
8. We will notify you that we can accept samples.
9. Bring samples by during the sample drop-off times posted on the Lab Calendar on the front page of the website.

Fine Print

▶ That’s it!
What Does It Cost?
The FGC and NGSC cores each have a price sheet posted under the Pricing button on the front page.

The FGC only serves clients in the IDOM/DRC, except for microarrays which are open to UPenn and other academic institutions.

In both cases, academics institutions outside of Penn pay a higher amount to make up for the lack of grant overhead.

For maximum flexibility there are separate charges for
1. sample and library quantification
2. library preparation
3. sequencing, charged by the lane
4. standard analyses, charged by the sample
5. advanced analyses (only provided by the FGC)

The exact cost of an experiment depends on many factors, so it is best estimated with the help of the FGC or NGSC technical director after submitting an experiment request.

How Does the Core Bill Me?
When you submit a New Experiment request to create or alter an investigation, we will contact your BA to make sure the budget code you have supplied is correct and can accommodate the charges we estimate at that time.

We do not bill at that time. We generate invoices roughly monthly that cover all work 'completed' in the billing period. Completed means that the bioA, sequencing run, analysis, etc. has a finish date in the billing period.

Thus it is possible to receive 3 invoices for a simple experiment if, for example, the bioA is done in October, sequencing is done in late November, and data processing is done in early December. Large, long experiments will usually generate a number of invoices.

We post the invoices on a password protected website (see URLs list). Normal PI and investigator logins work as well as a special set of BA accounts that allow BAs to see all of the PIs they handle.
Guidelines for Samples

Common Guidelines
1. Minimum 15µl or more
2. Please submit your samples in 1.5-2.0ml tubes
3. List of your samples by name on tube
4. Index number per sample
5. Source ID (Mouse ear tag #, cell line, etc.- something that ties this library back to its source)
6. Sequence length and type (50SR, 100SR, 100PE)

Please keep in mind that your samples will be handled often. In order to keep them from getting mixed up, please submit samples in tubes that are clearly labeled and intact (no unattached caps).

We will assign a sample id to each sample and attach a label with the sample ID, PI and investigator on each tube. Use this sample ID to track your samples.

Does the Core Make Libraries?
We offer library preparation services for a few different kinds of libraries. They are listed at the website along with the price. See the Prices button on the front page. Due to limited staffing we strongly encourage you to make your own libraries.

We can train you to make other libraries, but do not offer these services.

We are planning on purchasing a robot to automate library preparation, but for now library prep is a time-consuming task that can take a while for us to complete.

The library prep schedule is included under the Runs Calendar tab on the home page. You can check there to see when the first open slot is.

Sequencing Libraries
Here are our additional guidelines for submitting sequencing libraries.

1. Minimum 10nM concentration
2. Bring any custom primers if applicable

The concentration guideline is a recommendation for typical libraries. There are protocols that are known to generate lower concentrations. Include this information in the experiment request if you are aware of this situation.

RNA for RNA-Seq Libraries
Most library preps start with total RNA, reduction of ribosomal RNA is typically part of the library prep kit.

See posted protocols for typical RNA extraction methods.

1. 100ng/µl concentration
2. 10µl or more
How to Submit Samples

When Can I Bring My Samples In?
Do not bring samples in until we have told you that the investigation is ready for sample drop off. This means that we have an experiment description that we understand and have contacted your business administrator who has confirmed that you can afford the experiment.

If you already have an investigation and want to extend it to include new conditions or assays, then submit a new experiment request. We will process the request the next morning and once you hear from us you can bring the samples in.

If you are just bringing more samples for an existing condition/assay combination, you do not have to notify us before hand.

Finally, once we let you know we are ready for the samples, check the calendar to see when we have sample drop-off hours.

Shipping Samples to the Core
Samples should be shipped overnight to arrive Tuesday through Friday, i.e., do not ship over the weekend. These tips should help ensure the samples arrive in good shape.

1. Pack the samples in the center of a good amount of dry ice.
2. Place the tubes individually inside slightly larger tubes to ensure lids are not popped open during shipping
3. Clearly mark tubes
4. Include a paper print-out of the sample descriptions – see below
5. Email the core the tracking number as well as the sample description
Sample Descriptions

Overview

Using the description you submit with a New Experiment request, the core places samples in a hierarchy as shown below.

Investigation / Study / ( Condition × Assay )

For Condition and Assay we define terse tags which correspond to the experimental variables in your investigation.

Investigation usually contains the data for one, or maybe more related, papers.

Study is set of closely-related conditions and assays. Often one study may contain testing data, a second contains the bulk of the data, then the third study contains follow up data requested by reviewers.

Condition is the experimental condition you are manipulating, e.g., the day in a time series, cell type, genotype, treatment, or a combination of these.

Assay is not just RNA-Seq vs ChIP-Seq, but includes other details of the library technique, such as the antibody for a ChIP-Seq library. Basically it is roughly everything you do once you have the cells or tissue and begin to extract nucleic acids to start the process of making a library.

We have found this approach to reduce confusion about samples. Also doing this allows us to automate data analysis.
Reviewing Sample Descriptions

Reviewing the Experiment Design

To see the terms we are using take the following steps:

1. Click on the menu item Progress :: My Investigations (top of the page)
   - A page will appear that lists all of your PI's investigations. The investigation names at the left are the official names that the core uses.
2. Click on an investigation title to see details of the investigation.
   - A page will appear with a description of the investigation and one or more tabs for the studies under the investigation. You can cut and paste the study name into your sample spread sheet. Most investigations only have one study, so often this can be omitted.
3. Click on Study tab (if there are more than one studies) to see the conditions and assays.
   - Inside the study tab is a table of sample counts, i.e., how many samples of each condition and assay we have logged into the system.
   - Find the condition tag (top row) and assay tag (left column) that apply to your new samples and cut and paste them into your sample documentation.

Once you have samples checked in, the Progress :: My Samples page will list the samples and shows the Investigation, Condition, and Assay for each sample. This is another way of getting the exact text we will recognize easily when you submit new samples.

e-mail Alerts

Starting April 1, 2014 (no kidding!) the core has started send automated emails every evening to clients who have dropped off samples. Please check the annotation for the samples and let us know if there are any errors.
How Are My Samples Doing?

The Sample Queue

The sample queue currently covers only the steps from sample drop off to sequencing. It does not yet automatically handle resequencing, but we are working on that. The queue has the steps detailed below. This queue contains samples for both the NGSC and the FGC cores. The status of samples are updated every 15 minutes, but many of the initial steps occur in batches as we finish processing the various parts of the QC process.

Occasionally people can not find their sample in the queue. Here are some reasons why this can happen.

▶ when samples are pooled, they disappear from the queue and are replaced by the pool. This is normal.
▶ once a sample has been sequenced it will eventually disappear. This is normal.
▶ we accept a pool and mark is as 'have it in a pool'. This is incorrect - please let us know if we make this mistake!

Details of the queue are on the next page.

When Will My Samples Be Sequenced?

We have two areas of the website that display this information - the Queue (see above) and the Runs Calendar. The Runs Calendar shows runs from the past and present as well as runs that we are planning. Runs that are fixed will have names like FGC0600. The text will indicate if the run failed and how many lanes were present from each PI and investigator. For example PIGS/3 means that Paul Inigo's postdoc Grüvi Somers had 3 lanes' worth of samples. These are loaded from the database every few days. Runs that are planned will have human names, like #Elvis or #Eleanor. (These names are generated automatically and have no significance.) We will fill in samples numbers as we know them. These are subject to change but should firm up as their start date draws near. They will get pushed back when we have run failures or resequencing is needed.
Table of Contents

For People in a Big Hurry
   Quick Start
   How to Find Us

Just Enough Detail
   Sequencing at UPenn
   Getting Started
   Costs
   Bringing Samples
   Preparing Samples
   Documenting Samples
   How's It Going

Many Details
   The Website
   Microarrays
   *-Seq
   Sample Quality Checks
How Are My Samples Doing? - Details of Queue Status Tag

QC  Quality control involves performing at least an Agilent bioanalyzer run (for sizing) and a qubit measurement (for concentration.) Additional runs may be needed if these fail or give contradictory results. We are evaluating the use of the Kapa system for QC but this is still preliminary and is only being applied to samples selectively. We track the various evaluations in a spreadsheet, the upload values once all checks have been done and we are confident in the results. Thus in many cases work is being done on a sample that may not be reflected immediately in the queue. This step covers, RNA and DNA, libraries just being sequenced, as well as libraries that need to get resequenced due to bad sequencing results. After the bioanalyzer step, you will receive an email indicating the results, but further processing may be going to refine the concentration. After QC samples are either marked BAD and processed no further, or are marked GOOD and move to library prep, pooling, or ready for sequencing depending on the sample type.

Extraction  Extraction of RNA or DNA from cells is rare. Any RNA or DNA extracted will become a new sample and move to the QC queue.

Library Prep  Library prep covers making libraries from RNA or genomic DNA. These are typically time-consuming process that take away or two to process 8 samples. The libraries produced are new samples that move to the QC queue. Usually a bioA is done immediately, so he QC is to get the precise molarities.

Microarrays  Microarrays are performed by the FGC core. These typically take about 2 to 3 three person-days to perform. Due to this they may take a few weeks or more to schedule.

Pooling  Pooling covers the dilution and mixing steps necessary to sequence on or more libraries. We are very careful with this step as it is essential to achieving maximum read counts and even distribution across all libraries in a pool. We qubit the dilutions and redilute if necessary. Entries in this stage represent individual libraries which will be pooled to take up just a lane or two, usually.

Waiting to be Sequenced  This queue is the set of pools or individual samples that are ready to go. When a sample or pool will be sequenced in more than one lane, the queue entry will indicate this.

Sequencing Now  These samples are in runs that are going on ‘now’. This include runs that are just finishing or have recently finished. The end of a run may also be defined manually when a run is having trouble.

Recently Sequenced  Recently sequenced is just a list of the recent runs, in case you missed your sample going through the queue.
How Do I Download Data?

The Basics
There is a 'Download Data' on the website. This will lead to the download area for you/your PI's area. The site is password protected using your core username and password. All files that the Core produces for you will be available here.

Data is accessed via the http protocol (i.e., any web browser), but can easily be downloaded in bulk using command line utilities such as wget or curl. In addition, many GUI programs can also download from http. See the section on Data Files below to see how the files are organized.

Data Sharing
Our policy is to make all data for a PI available to all of her lab members.

Collaborators can get access to all of a PI’s data by setting up an account as a member of the PI’s lab (with permission of the PI of course!)

If a collaborator should only have access to a limited set of investigations, then the collaborating PI and relevant lab members should set up a set of standard PI and investigator accounts. We will then connect them to the appropriate investigations.

Bulk Downloads
If you have access to a command line, i.e., in Linux or Mac OS X, you can use the wget command as follows to, for example, download all of your FASTQ files (where you replace USER and PASSWORD with your username and password and PI and INVESTIGATION with your PI and the investigation of interest.)

```
wget --user=USER --password=PASSWORD -r -np \
http://fgc.genomics.upenn.edu/Experiments-2/PI/INVESTIGATION/basic/Fastq/
```

The -r -np options are very important and should not be left out.

- **-np** prevents wget from traveling up the hierarchy which can reach into the public data on the server (generating a huge download.)
- **-r** makes the download recursive which means that it will download nested folders.

At the moment we do not support ftp or Aspera connections for acquiring data.
Multiplexing allows you to sequence multiple samples in the same lane in order to minimize costs by controlling how many reads are allocated to each sample or to avoid confusing any lane-to-lane variation with differing experimental conditions.

Here are the important points to keep in mind:

- For small pools, the indexes must be chosen carefully to make sure they can be correctly decoded by the sequencer. Usually the library kit will provide rules for selecting a set of indexes.
- Reads are allocated between samples in a pool according to how each sample clusters compared to the others in the pool. This is controlled by how well we are able to measure the effective molarity of each sample as well as our pipetting skills. The number of reads will always vary somewhat from the ideal.

From an experiment design point of view, there are occasionally small lane-to-lane variations in the inherent biases in sequencing. These are usually minor, but to remove any doubts about this problem, it is best to avoid putting all libraries from one condition in a single lane or pool. Rather spread the libraries from each condition across multiple lanes or pools. In the most extreme case, make a large pool of all libraries and sequence this pool across enough lanes to achieve the needed coverage.
What On-Line Tools Can I Use to Analyze My Data?

Here is a short list of on-line tools for genome data analysis and/or visualization:


**HOMER** ChIP-Seq and other data analysis - [http://biowhat.ucsd.edu/homer/ngs/index.html](http://biowhat.ucsd.edu/homer/ngs/index.html)

**CISTROME** ChIP-Seq analysis - [http://cistrome.org/Cistrome/Cistrome_Project.html](http://cistrome.org/Cistrome/Cistrome_Project.html)

**GALAXY** ChIP-Seq and other analysis and visualization - [https://main.g2.bx.psu.edu](https://main.g2.bx.psu.edu)

**UCSC Genome Browser** - [http://genome.ucsc.edu](http://genome.ucsc.edu)
Many Details

Microarrays

Microarray Experiments

Introduction
The FGC core offers microarray services. We use the Agilent platform. The arrays are spotted 4 arrays per slide. We usually do two-color experiments for most direct comparison between two conditions. For multi-condition experiments we generally do single-color experiments. There is a standard pipeline for each of these two types of experiments.

What Arrays Do We Offer?
The standard microarrays we use are the Gene Expression mouse or human 4x44K. Other species or 8x60k arrays are can be special ordered, but this takes up to two months for delivery.

1-Color Experiments
In this class of designs each sample is labeled with Cy3 (green) dye an hybridized to its own array. This enables comparisons between any pair of conditions but at the cost of slightly lower sensitivity and higher expense for more arrays.

2-Color Experiments
In this class of designs, samples are paired and within each pair one sample is labeled with Cy5 (red) and the other with Cy3 (green) and cohybridized on the same array. The analysis is done with the ratio of sample pairs rather than with individual samples. This design has the benefit of reducing contamination effects when samples are properly paired as well as reducing costs, but limits the comparisons that can be performed.

How Many Samples Do I Need?
We strongly recommend at least 4 biological replicates. This means that you have decent statistical power, robustness to hybridization failures, and are utilizing the arrays efficiently.

How to Bring the RNA
For the normal microarray kit we need 10ul of RNA with a concentration of at approximately 100ng/ul. Do not submit very concentrated RNA as this will add processing time to the experiment. We will return unused RNA. If you can not get this much RNA, we can amplify the RNA (at extra cost). Then the requirements are 20-50ng of total RNA in up to 5ul volume. We use the NuGEN Ovation Pico kit used for amplification prior to labeling and hybridization. We check the quality and quantity of the RNA on an Agilent bioAnalyzer. We strongly suggest that the RNA quality be high with a RNA integrity number (RIN) of 8 or more.
Replicates

In general we recommend 2 or more biological replicates whenever possible. More is better if you can acquire the samples and can afford the sequencing costs.

With 2 replicates we may be able to calculate a p-value but if there are discrepancies between samples, you may need a 3rd replicate to identify the outlier.

P-values from 1 replicate are very unreliable (when they can be calculated) but a single replicate can provide a starting point for validation at the bench.

Depth of Sequencing

The depth of sequencing depends on the type of library and the experimental question. Here are our starting points:

<table>
<thead>
<tr>
<th>Library Type</th>
<th>Assumptions</th>
<th>Samples per hiSeq Lane</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-Seq</td>
<td>microarray-like power</td>
<td>about 4-8</td>
</tr>
<tr>
<td></td>
<td>more power</td>
<td>2 to 4</td>
</tr>
<tr>
<td></td>
<td>deep sequencing for rare splicing events</td>
<td>1 or less</td>
</tr>
<tr>
<td>ChIP-Seq</td>
<td>transcription factor in a mammal</td>
<td>about 6</td>
</tr>
<tr>
<td></td>
<td>weak antibody or diffuse binding pattern</td>
<td>1 to 3</td>
</tr>
<tr>
<td></td>
<td>strong reference input</td>
<td>1</td>
</tr>
<tr>
<td>miRNA-Seq</td>
<td>10 million reads per sample.</td>
<td>15 to 20</td>
</tr>
</tbody>
</table>
Library Construction

Library construction is necessary to adapt your RNA or DNA so that the Illumina Sequencers can manipulate the material to sequence it. This basically amounts to adding specific adapter sequences to each end of 'small' DNA fragments. This process differs greatly depending on the experiment but the end results look fairly similar.
Details to follow ...

Multiplexing

See the Multiplexing in the First Things First section.
RNA-Seq

Introduction
There is a wide variety of kits available for making RNA-Seq. In selecting one there are a few basic questions that will help sort out which ones to consider.

1. How do you want to reduce the amount of ribosomal RNA - polyA selection or ribosomal depletion?
2. Do you want to preserve the direction of transcription?
3. How much total RNA can you get per sample?

Ribosomal Reduction
Since ribosomal RNA constitutes about 90% of total RNA, it has to be removed somehow, otherwise most of your reads will be ribosomal and will contribute nothing to the experiment. The two main methods polyA-selection and ribosomal removal are both effective, but ribosomal removal can be more expensive although essential if you are interested in capturing non-coding RNAs.

Preserving the Sense of Transcription
In many cDNA-based library preparation methods, double-stranded cDNA is made from the RNA, then used to make a library following the standard Y-shaped adapter Illumina methods. In this scenario both strands of the cDNA can be sequenced and there will be no direct way to tell whether a read comes from the sense or anti-sense strand. This may not be a major problem, but there are times when the sense helps disambiguate read assignment with overlapping genes. I can also help determine the structure of novel genes, which these days are mostly non-coding RNAs (in human and model organisms.)

Amount of Starting Material
The standard Illumina truSeq RNA-Seq kits work very well with about 200ng of total RNA. The user’s guide indicates 100ng can work, and with some luck you may be able to push the amount even lower. Once you get much lower, say 25ng, you will want to switch to a different kit. The NGSC/FGC core often uses the NuGEN Ovation PicoKit to amplify small amounts of RNA. (This kit does the ribosomal depletion as a part of the amplification process.)
ChIP-Seq

Measuring Enrichment Prior to Library Prep

Test your antibody and sample prep before making a library. We recommend that the enrichment ratio \( \frac{C^+}{I^+} \geq 10 \). Where \( C^+ \) is ChIP at positive control, \( I^+ \) is input at positive control, \( C^- \) is ChIP at negative control, and \( I^- \) is input at negative control.

To do this, you need at least two primer pairs, a positive control (+) and a negative control (-) which you measure on the ChIP and input samples using Q RT-PCR. See the figure to the right.

Control

We strongly recommend that you sequence an input library for each condition. Note that in the figure, the input track has a strong peak at the same place as the ChIP peak. The strength of inputs peaks or bias is dependent on the state of the cells as well as chromatin preparation conditions.

Sequencing

- ChIP-seq libraries are normally sequenced with 50bp SR runs on the hiSeq.
- ChIP-seq is primary a counting process, the reads only have to be long enough to place most of them uniquely in the genome
- 40 million reads are usually sufficient in a mammal.
- We have a multiplexed library protocol which can be used to put about 6 libraries in one lane.
- You may need more reads if the target protein is spread across large sections of the genome.

Sonication

Chromatin fragmentation is accomplished either via sonication or (for histones) DNase treatment. The fragmentation conditions can dramatically affect the results, so be consistent.

- Sonication needs to be tuned to the sample.
- Successful sonication is a balancing act between a few opposing trends listed below.
- Some complexes are fragile so use a little sonication as possible.
- We only sequence the fragments with lengths of about 150bp, so enrichment in long fragments does not help.
- More sonication reaches deeper into dense chromatin.
- For low cell counts, you need to be as efficient as possible, so you need to leave as little material outside the fragment lengths that actually get sequenced.
- With the proper primers, you can verify enrichment in both the ChIPed chromatin and the library to ensure that the enrichment is still present in the library.
Interpreting BioAnalyzer or TapeStation Traces

Introduction
The cores use the Agilent bioAnalyzer and/or TapeStation to get the length distribution of DNA fragments in a library. These are sent to clients before proceeding with sequencing. This section provides a brief description of what the trace should look like for a variety of different library types.

Most modern libraries use the Illumina truSeq adapters which are 120bp in total length. Thus the sequenceable insert is 120bp less than the annotated peak sizes in a trace.

General Points
Check the markers to make sure that the software identified the short and long ladders steps correctly. If the ladders are mis-identified, then the calculated library lengths with be incorrect.

By default the bioA software identifies sharp peaks, so we often have to set up a manual integration region. In a ChIP-seq library for example, it may split the big obvious peak in to a two or more sections based on very tiny wiggles across the top of the peak.

RNA-Seq
RNA-Seq libs are similar to ChIP-Seq, but may exhibit a long tail toward longer insert sizes. The distribution should be smooth.

miRNA
Since miRNAs have a very limited range of lengths and the library construction protocol usually involves a strong size selection step, miRNA-Seq libraries usually have a very sharp peak at about 142bp with a small shoulder, perhaps up to 150bp.

Exome-Capture

Amplicon Targeted Sequencing
An amplicon-based targeted sequencing library can produce either a peaked distribution with one or a few abundant lengths, a relatively smooth distribution, or something somewhere in between. This depends on the number and distribution of sizes of the amplicons and the application. A single amplicon from one location of the genome would normally produce a single peak. If you are using degenerate primers, e.g., to pick out repeat inserts, then your amplicons will be coming from multiple locations in the genome and will probably have a range of lengths. On the other hand, a relatively large panel of genes will probably produce a smoother distribution as the positions of PCR primer pairs are adjusted to obtain best performance in the local CG and repeat content of the target genes. It is best to understand your library prep technique well when assessing a gel trace.
Common Problems

Problems with samples or libraries are often caught before sequencing by our QC steps, but others are not apparent until sequencing has been done. Here we walk through the common problems in the order that they are usually detected.

Adapter Dimers
The cores usually analyze libraries with either an Agilent bioanalyzer or tapestation. In need more here!

High Read Redundancy
Most library preparation protocols there is a PCR step which amplifies the library and possibly extends adapters to the full complete sequences needed.

Bacterial Contamination
The Problem A ChIP-seq, RNA-Seq, or other library type looks good and sequences well, but has a very alignment percentage, i.e., much less than 50%. Sometimes this is due to poor library construction that results in artefactual sequences, but once in a while the library is good but contains a large amount of bacterial sequence and relatively little of the intended species.

Confirming the Diagnosis In the blastn section of the NCBI BLAST website (http://www.ncbi.nlm.nih.gov/blast) there is a 'Whole-genome shotgun contigs (wgs)' database which contains sequence from a wide range of organisms. If you convert 20 or so of your reads to FASTA format, you can paste them into the search window and see what species they match. If you get a bunch of excellent matches to a species (other than the one you had hoped for) then that’s the problem. The exact species can depend on the source of the contamination, i.e., local water, reagents, or from the bacteria in the host species, e.g., gut flora/fauna or dirt etc.

Tracking Down the Source For bacterial contamination the easiest way to find the source is to do PCR on water samples to see if you can find ribosomal sequence. The primers below hit the 16rDNA gene in a wide variety of species but do not match mammals.

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward primer</td>
<td>5’-TCCTACGGAGGCAGCAGT-3’</td>
</tr>
<tr>
<td>Reverse primer</td>
<td>5’-GGACTACCCAGGGTATCTAATCTTGTT-3’</td>
</tr>
<tr>
<td>Probe</td>
<td>(6-FAM)-5’-CGTATTACCGGGCTGCTGGC-3’</td>
</tr>
</tbody>
</table>

Something’s Fishy!
Another possible problem is the use of salmon DNA as a blocking agent in ChIP-Seq experiments. Adding a blocking agent is harmless when probing a few locations with PCR. However, the salmon DNA ends up in the library. This is rare these days as the cores routinely warn against it and the issue is more widely reported. The blast search above should identify this problem as well.
Downloading Data

This section describes the downloads section of the core website. In addition it gives a brief introduction to common file formats, limited to what kind of information they contain.
Accessing Data via Website

The Download Data Button

The Download Data button on the website leads to web pages that provide direct access to all of the files for your PI. You will have to login again using your normal core account username and password. Inside this directory will be one or more folders that correspond to the investigations your PI has us. Inside each investigation folder are another set of files and folders which are fairly standardized.

AAA-StudyInfo.xls/.txt these files describe the sample and any sequencing we have done for this experiment. They are update automatically.

Manual this folder contains our notes for the investigation as well as custom programs and small sets of data.

basic this folder contains the raw and lightly processed data for the investigation. More details below, but basically this is where the FASTQ files, any trimmed FASTQ files, alignment files, and simple track files will be.

Analysis this folder contains more advanced standard and custom analyses. For example peak calls and motif finding results will be here. More details below in the various Pipeline sections of this document.

Where’s My Investigation?

There are two main reasons why an investigation may not appear when you click the Download Data button. First, it’s an old investigation, or it is part of a collaboration with another PI.

If the investigation is old, try making the following change to the URL in your browser. Instead of http://fgc.genomics.upenn.edu/Experiments-2/PI replace the Experiments-2 with Experiments-1. Also bear in mind that some very old experiments are named using the invoice number, PI, and investigator, so you may have to pick through a few folders to find the one you want.

If you are a collaborator and need to find investigations from other PIs, try replacing your PI’s name in the URL with the name of PI of the investigation.

If either of these do not work, please feel free to contact us.
The basic Folder

The basic folder contains files that hold the basic ‘raw’ sequence data, any trimming of adapter and/or low quality sequence as well as the alignments to one or more genomes. It also contains the visualization data that appears in the TessLA browser for this data, e.g., the USHP tracks. The contents of this directory have evolved over the time the core has been open, so older experiments may have a slightly different set of files that new ones.

Current Practice

FASTQ  these are the ‘raw’ data from the sequencer.
  ▶ Raw is in quotes as the actual output of the sequencer is in BCL format which we convert to FASTQ.
  ▶ The files are usually (but not always) compressed with gzip.
  ▶ FASTQ files will be uncompressed during the initial analysis then compressed again later on.
  ▶ The files are named by the RUN, LANE, END (for paired-end sequencing) and BARCODE of the data. For example FGC0503_s_1_1_AGGCAGAA.fastq.gz is the data for run FGC0503, lane 1, end 1, and barcode AGGCAGAA.
  ▶ You may occasionally see 'Undetermined' or 'OTHER' instead of a barcode. This happens when we either do not know the barcodes or there is a discrepancy between the reported barcodes and what we see in the pool. When OTHER is present, we have probably split the Undetermined file using the observed barcodes.

fastqc  Often we run the utility FASTQC to assess the basic statistics of the sequenced generated to look for adapter, base qualities, CG bias, repeated sequence etc. The output of this program consists of both an HTML report and other pictures and text files.

Bowtie  output of bowtie aligner including alignments as well as logs and statistics of the alignments.

BedFiles  alignments converted to BED or other similar formats for visualization in the TessLA browser.

Previous Practice

Solexa  This folder contains the sequencing (sequence.txt) and alignment data (export.txt) sequence.txt files can be converted to FASTQ for data submission or updating analyses.

Config  can generally be ignored, contains meta data about sequencing and samples.

Export  contains processed alignment data.
The Analysis Folder

The UCSC genome browser can load and generate a number of different file types. Their website has a good description of the formats, but we can add a few comments here.

**BED Files**
BED files are the standard method for delivering locations of 'small' sets of feature, e.g., 40,000 transcription factor binding sites. The files do not contain any sequence information.
BED file can be uploaded to UCSC and many other genome browsers to view results.

**Fastq Files**
Fastq are the standard format for delivering 'raw' sequencing results.
Fastq file contain a unique ID for each read, the read sequence, and base qualities.
The files do not contain any alignment information.
Standard Pipelines

The FGC and the NGSC offer a reasonable set of standard processing pipelines for most types of sequencing experiments. The FGC can offer customized analysis to the extent that we have the staff hours to accommodate requests, but is only open to DRC/IDOM members.
The following few slides describe the standard pipelines that we offer.
Regional Enrichment - aka Peak Calling

Introduction This tool is used to identify statistically significant enrichment on regions that are defined relative to annotated regions of the genome, rather than to regions defined by the ChIP-Seq data itself. It is typically used when the pattern of the input ChIP-Seq target is so diffuse that standard peak callers have a difficult time identifying regions of enrichment. In this case we will use regions that are of a priori interest, such as promoters, gene bodies, CpG islands, etc. that are likely regions to contain enrichment for the ChIP target. The ngsc-chipseq-RegionalEnrichment tool counts reads on these pre-defined regions of interest for both a ChIP and a control (usually input) sample, then uses a Fisher exact test and Benjamini-Hochberg correction to assess the enrichment of the ChIP signal on the region. A new track will be loaded with the enrichment ratio as the score and the p-value and FDR.

Details A pseudocount of 1 is added to each region when computing the Fisher test.
ConnectSpanToGene

Introduction
We routinely run the ConnectSpanToGene tool to associate ChIP-Seq peaks (or many other types of regions) with genes, but it can be run to associate any set of regions with genes.

ConnectSpanToGene takes a gene track, a span track (which has the peaks), and parameters (MaxDistBp, ToleranceBp, TolerancePct) controlling how far away from a span we will look for a gene. It outputs a tab-delimited file containing the results, which we usually convert to an Excel file and load into the database (attached to the region track).

How Does it Work?
The program considers each region in the region track. It then reports any gene that overlaps with the span. It then puts the nearby genes in order by the distance from the span to the transcription start site (TSS), with the closest TSS first. If the distance (D) to the first gene is less than the distance threshold (MaxDistBp), then the gene is reported. Any other genes that are closer to the span than \( D \times (1 + \frac{TolerancePct}{100}) \) and \( (D + ToleranceBp) \) are also reported. This process is then repeated for each span in the span track. The value of D is adjusted for each span.

How do We Usually Run it?
We usually run with the following settings:
MaxDistBp = 100,000

What Does the Output Look Like?
Here are the columns in the output file.

<table>
<thead>
<tr>
<th>Column</th>
<th>Values</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Span-GenomeRelease</td>
<td>e.g., mm9</td>
<td>genome of the spans (and the genes)</td>
</tr>
<tr>
<td>Span-Chromosome</td>
<td>e.g., chr1</td>
<td>chromosome of spans</td>
</tr>
<tr>
<td>Span-BeginBp</td>
<td>integer</td>
<td>begin of span</td>
</tr>
<tr>
<td>Span-EndBp</td>
<td>integer</td>
<td>end of span</td>
</tr>
<tr>
<td>Span-Strand</td>
<td>'+' or '-'</td>
<td></td>
</tr>
<tr>
<td>Span-Score</td>
<td>float</td>
<td>the meaning of the value depends on the span table.</td>
</tr>
<tr>
<td>Span-Name</td>
<td>string</td>
<td></td>
</tr>
<tr>
<td>GeneI</td>
<td>empty, or 1, 2, 3 etc.</td>
<td></td>
</tr>
<tr>
<td>AbsDistanceBp</td>
<td>integer</td>
<td>absolute value of distance from span to gene TSS</td>
</tr>
<tr>
<td>DistanceBp</td>
<td>integer</td>
<td>distance from span to gene TSS. Positive values are downstream of the TSS.</td>
</tr>
<tr>
<td>Overlaps</td>
<td>yes/no</td>
<td>does the span overlap the gene.</td>
</tr>
<tr>
<td>Gene-GenomeRelease</td>
<td>e.g., mm9</td>
<td>genome of the gene</td>
</tr>
<tr>
<td>Gene-Chromosome</td>
<td>e.g., chr1</td>
<td>chromosome of the gene</td>
</tr>
<tr>
<td>Gene-BeginBp</td>
<td>integer</td>
<td>beginning of the gene</td>
</tr>
<tr>
<td>Gene-EndBp</td>
<td>integer</td>
<td>end of the gene</td>
</tr>
<tr>
<td>Gene-Strand</td>
<td>'+' or '-'</td>
<td></td>
</tr>
<tr>
<td>Gene-Score</td>
<td>float</td>
<td>Usually meaningless, but could be a differential expression value.</td>
</tr>
<tr>
<td>Gene-Name</td>
<td>string</td>
<td>name of gene</td>
</tr>
<tr>
<td>Span-Pvalue</td>
<td>empty or float</td>
<td>p-value for detection of the span</td>
</tr>
<tr>
<td>Span-FDR</td>
<td>empty or float</td>
<td>FDR for detection of the span</td>
</tr>
<tr>
<td>Span-ContentTag</td>
<td>empty or text</td>
<td>extra stuff about the span. Varies with the span table.</td>
</tr>
</tbody>
</table>
Like most experiments, we initially process RNA-Seq samples individually in a run/lane/barcode (or RLB) phase. We use RUM for this step. Next we proceed to call differentially expressed genes using a custom pipeline described below.

Cleaning
We first align reads to ribosomal sequences and repeats remove these reads and to assess the level of ribosomal sequence in the libraries.

RUM
We use the RUM package from Grant et al to do the basic processing of RNA-Seq data. RUM generates a set of files which we then process a bit further to make them visible in the TessLA browser and for other down-stream analyses.

Files
Here is a typical set of files produced for a RUM analysis with typical files sizes in bytes.
Comparison

Summary
We run this tool to do basic differential analysis. It is best used for RNA-Seq data, but can be used for other data as well.

Files
By default the files are created in Analysis/DiffExp. In this directory, you may find multiple analyses which use different data and/or parameters. Looking inside one of these directories, you will see 3 to 4 files with names like Compare.*, the most useful of which is Compare.tab.xls.

Compare.tab.xls
This file contains the comparison data. The contents are somewhat flexible, but will follow this outline. Each row is a transcript. The first few columns contain the gene, transcript, and 'Best' (an indicator which guides you to the best transcript for each genes.) The next set of columns are various comparisons. Which comparisons are done depend on the experiment. For each comparison there are 6 columns.

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Data in the Column</th>
</tr>
</thead>
<tbody>
<tr>
<td>MVA:M:Test:Control</td>
<td>log2 test/control fold change</td>
</tr>
<tr>
<td>MVA:A:Test:Control</td>
<td>log2 average expression</td>
</tr>
<tr>
<td>EDGE:A:Test:Control</td>
<td>log2 average expression</td>
</tr>
<tr>
<td>EDGE:M:Test:Control</td>
<td>log2 test/control fold change</td>
</tr>
<tr>
<td>EDGE:pv:Test:Control</td>
<td>0-1 p-value</td>
</tr>
<tr>
<td>EDGE:FDR:Test:Control</td>
<td>0-1 FDR from p-value using Benjamini-Hochberg correction</td>
</tr>
</tbody>
</table>

- The first word in each column title indicates the tool that is used to produce the data in the column.
  - MVA is a simple MvA comparison with no statistical significance.
  - EDGE is the EdgeR package which performs differential gene expression on RNA-Seq data.
  - The data that is passed to the analysis programs has been quantile normalized.
- The second word indicates what type of data is in the column.
  - M values are the log2(Test/Control), so M=1 indicates 2-fold increase in expression.
  - A values are log2 of the average expression between two conditions. MvA and EdgeR use different units, MVA is usually Reads, whereas EdgeR values have been normalized to counts per million.
- The ‘Test’ and ‘Control’ words are replaced with the actual conditions used in the comparision, e.g. ‘KO’ and ‘WT’.
- The next set of columns of the file are quantile normalized log2 versions of the 'raw' data for the individual samples.
- The last set of columns are the 'raw' data which is usually reads.

Looking Deeper
Within each Comparison folder is another called 'Heatmap'. See http://fgc-ngsc-cores.blogspot.com/2012/09/analysis-tools-rum-multiplecomparisons.html for details about the files in this folder.
RUM-MultipleComparisons

Introduction
We routinely run the pipeline RUM-MultipleComparisons to assess RNA-Seq data. Although the tool includes the work 'RUM' in the title, it can work with gene expression values from a variety of RNA-Seq tools. We are still expanding what analyses RUM-MultipleComparisons performs but at the moment, it includes these basic steps.

1. Assemble a table of the raw data
2. Filter to consider just transcripts
3. Performs quantile normalization of the values
4. Performs k-means clustering of the data and displays results as heatmaps
5. Generates MvA plots of averages for all conditions
6. Generates MvA plots of replicates within a condition
7. Tabulates fold-changes between average values for all conditions

What Files Should I Look At?
First, take a look at the plot, Replicates and Kmeans-heatmap.pdf files so that you can see if the samples have good intra-condition consistency. In addition, the heatmap file will help you see if the changes between conditions are consistent across samples, and roughly how many sets of expression patterns there are in the set. Once you can see that the data is ok, turn to the Averages.tab file or the appropriate Kmeans-*.clusters.tab file to see gene IDs. All of the tab files can be opened from within Excel which can be used to further filter the genes. Gene lists can also be created for use with functional analysis.

How Do we Usually Run It?
We usually focus on well-characterized RefSeqs, i.e., those with IDs like NM_* or NR_*.

What Does the Output Look Like?
Plots
1. AllPairs-mva.png - a comparison of all samples in the data set.
2. Kmeans-heatmap.pdf - series of heatmaps using different numbers of clusters. Yellow/white is high expression, red is low.
3. Pairs.pdf - MvA plots of all condition comparisons
4. Replicates-mva.pdf - MvA plots of replicates within a condition
5. Tables of Data
6. AllTranscriptReadCounts-sql.tab - initial raw data
7. AllTranscriptReadCounts.tab - data filtered to just transcripts
8. Averages.tab - averages over conditions with fold-changes for all comparison
9. Details-Lg2-Qn.tab - quantile normalized values for individual samples
10. Kmeans-04-clusters.tab - details of genes in each cluster.
11. Kmeans-05-clusters.tab
12. Kmeans-06-clusters.tab
13. Kmeans-28-clusters.tab
14. Kmeans-29-clusters.tab
15. Kmeans-30-clusters.tab
...