An Introduction to NGS at UT: The UT GSAF

Scott Hunicke-Smith
Updated: 9/23/14
Where to find more information

gsaf.cssb.utexas.edu

Ready to submit samples? Use these links

7/25/13 NOTE: UT's new version of Confluence requires you to "Log In" in the upper right corner BEFORE selecting an online submission form link.

To submit a new project or get detailed pricing information, use these links. You'll be given a 6-digit request ID to refer to your request later. Your requests stay here until you're ready to submit samples. You return to this link when you're ready to submit your samples.

1. Log in in the upper right corner with your EID (don't have an EID, or haven't submitted a sample to us before? Please go read this page).

2. Choose your project type:
   1. Bacterial metagenomics project - Simplified submission page for bacterial 16S metagenomics (you must log in first - upper right corner)
      - Metagenomics assay: This service starts with normalized (i.e. equal concentration) DNA samples and includes amplification of the V4/V5 regions of the bacterial 16S gene as shown on this web page and at least 10,000 2x250 bp paired-end sequences from the Illumina MiSeq platform for at least 95% of the samples submitted. PCR is performed in triplicate to minimize jackpot effects. Gel-based QC is performed on a sampling of 10% of the samples post amplification. DNA is expected as input; the GS AF does not provide DNA extraction services at this time. See this web page for sample input guidelines and note that DNA concentrations should be normalized before submission. (The GS AF can normalize your sample concentrations for an additional charge.)
   2. All other projects (you must log in first - upper right corner)
      - Download sample submission template (required for jobs with more than 30 samples; optional otherwise) for the general, non-bacterial metagenomics submission page ONLY: Projects must be submitted via the online project submission form above to specify project details like platform, contact info, etc., but if you'd rather fill out an Excel template for your sample info instead of the web form, use this template. You'll upload it when you request a new sequencing job. This form is required for more than 30 samples; suggested for more than 10.

Latest news

11 Sep 14
Scott Hunicke-Smith presented GS AF updates at the Byte Club meeting. The presentation covers new user-accessible equipment, the NextSeq 500 and HiSeq V4 upgrade, pricing changes, new library prep methods, and new data storage options. Here is a PDF copy of the presentation.
What the GSAF does

• Sample QC
  – By Fluorimetry/Qubit and BioAnalyzer
  – Self-service and full-service

• Library Preparation
  – Construction of NGS sequencing libraries from DNA and RNA of all sorts

• Sequencing
  – Final sequencing library quantification
  – Operation of the Illumina HiSeq 2500, NextSeq, and MiSeq systems

• Advise and consult
  – On experimental designs, lab methods, and cost
What the GSAF does

• Sample QC
  – By Fluorimetry/Qubit and BioAnalyzer
  – Self-service and full-service

• Library Preparation
  – Construction of NGS sequencing libraries from DNA and RNA of all sorts

• Sequencing
  – Final sequencing library quantitation
  – Operation of the Illumina HiSeq 2500, NextSeq, and MiSeq systems

• Advise and consult
  – On experimental designs, lab methods, and cost

850 BioAnalyzer runs/year
273 384-well qPCR plates/year
7,500 samples/year,
190 sequencer runs/year
800 jobs (projects)/year
What the GSAF does not do

• Data analysis
  – See the [CSSB/CCBB Bioinformatics Consulting Group](http://) or
  – Investigate [data analysis training available at UT](http://) (both via CCBB)

• Design your experiment
  – You, the customer, are responsible for your experimental design including the detailed choices of library preparation, sequencer run type, etc.
  – Illumina has some excellent tools and training videos that can assist when planning out projects, please visit [http://support.illumina.com/](http://support.illumina.com/) there is also an excellent resource for calculating coverage at [http://support.illumina.com/downloads/sequencing_coverage_calculator.html](http://support.illumina.com/downloads/sequencing_coverage_calculator.html)
    • The coverage calculator will be extremely helpful when submitting your samples for sequencing and you need to determine the number of reads to request for each sample
How do I start a project with the GSAF?
How NGS sequencers work

• Please see this YouTube video for a great illustration:
  – [http://www.youtube.com/watch?v=77r5p8IBwJk](http://www.youtube.com/watch?v=77r5p8IBwJk)

• Important practical conclusions:
  – Sequencing library DNA must be properly formed and suited to your question
  – The quality and quantity of sequencing library is essential
### Sequencer run types & price

<table>
<thead>
<tr>
<th>Run Type</th>
<th>2014 rates</th>
<th>2013 rates</th>
<th>$/Mbp</th>
<th>$/(M reads)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HiSeq 2500 H.O. v3</strong></td>
<td>1x50</td>
<td>N/A</td>
<td>$955</td>
<td>$0.12</td>
</tr>
<tr>
<td></td>
<td>1x100</td>
<td>N/A</td>
<td>$1,318</td>
<td>$0.08</td>
</tr>
<tr>
<td></td>
<td>2x50</td>
<td>N/A</td>
<td>$1,656</td>
<td>$0.10</td>
</tr>
<tr>
<td></td>
<td>2x100</td>
<td>N/A</td>
<td>$2,317</td>
<td>$0.07</td>
</tr>
<tr>
<td><strong>HiSeq 2500 H.O. v4</strong></td>
<td>1x50</td>
<td>$1,013</td>
<td>N/A</td>
<td>$0.09</td>
</tr>
<tr>
<td></td>
<td>2x75</td>
<td>$2,091</td>
<td>N/A</td>
<td>$0.06</td>
</tr>
<tr>
<td></td>
<td>2x100</td>
<td>$2,428</td>
<td>N/A</td>
<td>$0.06</td>
</tr>
<tr>
<td></td>
<td>2x125</td>
<td>$2,766</td>
<td>N/A</td>
<td>$0.05</td>
</tr>
<tr>
<td><strong>NextSeq 500 H.O.</strong></td>
<td>1x75</td>
<td>$1,999</td>
<td>N/A</td>
<td>$0.08</td>
</tr>
<tr>
<td></td>
<td>2x75</td>
<td>$3,329</td>
<td>N/A</td>
<td>$0.07</td>
</tr>
<tr>
<td></td>
<td>2x150</td>
<td>$5,025</td>
<td>N/A</td>
<td>$0.05</td>
</tr>
<tr>
<td><strong>MiSeq V3</strong></td>
<td>2x300</td>
<td>$1,813</td>
<td>$1,720</td>
<td>$0.14</td>
</tr>
<tr>
<td><strong>MiSeq V2</strong></td>
<td>2x250</td>
<td>$1,396</td>
<td>$1,321</td>
<td>$0.21</td>
</tr>
</tbody>
</table>

- All run types indicate (# of reads) x (read length)  
  - e.g. “1x50” means single-end read, 50 bp long  
- Any run type can be single- or dual-indexed
Sequencer run types & price

- Requested Platform (will be set automatically for Metagenomics or ddRAD)
- Requested RunType (will be set automatically for Metagenomics or ddRAD)
- Buy sequencing by: (will be set automatically for Metagenomics or ddRAD)

Special Sequencing Notes: ONLY for comments pertaining to running the Illumina Sequencer, lane layout, etc. DO NOT put LIBRARY PREPARATION instructions here.

Buying per lane? Select "Fraction of Plate", then for HiSeq 2500 High Output lanes, enter 12.5% for each lane, for HiSeq 2500 Rapid lanes, enter 50% for each lane, and for MiSeq lanes - one run is one lane (i.e. 100%).

- All run types indicate (# of reads) x (read length)
  - e.g. “1x50” means single-end read, 50 bp long
- Any run type can be single- or dual-indexed
Sequencer run types & price

- All run types indicate (# of reads) x (read length)
  - e.g. “1x50” means single-end read, 50 bp long
- Any run type can be single- or dual-indexed
Sequencer layouts & capacity

• Data capacity per run:
  – HiSeq: 8 lanes/run, 220 million reads per lane
  – NextSeq: 1 lane/run, 330 million reads per lane
  – MiSeq: 1 lane/run, 13 or 22 million reads per lane

• Barcode capacity per run:
  – Single-indexes: typically MAX of 12 per lane, unless prepared by the GSAF who have >48 barcodes available
  – Dual-indexes: typically >12, up to 400 per lane
  – Single- and dual-indexes can be run in the same lane IF they can be separated after the run
Purchasing “by read”

• The UT GSAF is known for selling “by the read” and not just “by the lane”

• What this means:
  – Target vs. Minimum:
    • We aim to sequence “Target”
    • If we fall below Minimum, we will automatically re-queue your sample for more sequencing
    • You will be billed for the LESSER of Target or what you actually receive
  
  – Minimum order per instrument:
    • HiSeq: 20 million read-pairs
    • MiSeq: 1 million read-pairs
    • NextSeq: > of 20 million reads per sample or 40 million reads per job
Purchasing “by read”

“Fraction of Plate” allows you to enter a % below

Buy sequencing by: (will be set automatically for Metagenomics or ddRAD)

Special Sequencing Notes: ONLY for comments pertaining to running the illumina Sequencer, lane layout, etc. DO NOT put LIBRARY PREPARATION instructions here.

Buying per lane? Select "Fraction of Plate", then for:
HiSeq 2500 High Output lanes, enter 12.5% for each lane
MiSeq lanes - one run is one lane (i.e. 100%).
NextSeq 500 lanes - one run is one lane (i.e. 100%).

Fraction of Plate for Job

Enter Sample Information *
Purchasing “by read”

“Number of Reads” allows you to enter a number PER-SAMPLE on a later screen or via Excel file upload.

Buy sequencing by: (will be set automatically for Metagenomics or ddRAD) *

Special Sequencing Notes: ONLY for comments pertaining to running the Illumina Sequencer, lane layout, etc. DO NOT put LIBRARY PREPARATION instructions here.

Buying per lane? Select “Fraction of Plate”, then for:
HiSeq 2500 High Output lanes, enter 12.5% for each lane
MiSeq lanes - one run is one lane (i.e. 100%).
NextSeq 500 lanes - one run is one lane (i.e. 100%).

Enter Sample Information *

○ Manually
○ Attachment
Library Types

• DNA
  – Whole genome
  – Exome
  – ChIP
  – Amplicon
  – RAD

• RNA
  – Whole transcriptome – poly-A(+) or Ribosomal(-)
  – Small RNA
  – RIP
# Library Types

## DNA – listing on LIMS

- **Library Prep Required**
  - Yes
  - No

Library Types: NOTE - if you require special library prep handling (at extra cost), you must select "Special Fragmentation" and enter details under the Library Type Specifications field

- Standard DNA Library
- Mate-Pair Library
- Exome or Custom Capture
- Special Fragmentation
- DNA low cost high throughput
- ddRAD genotyping
- ddRAD development phase 1
- ddRAD development phase 2
- 16S Metagenomics

## RNA – listing on LIMS

- **Library Prep Required**
  - Yes
  - No

Library Types: NOTE - if you require special library prep handling (at extra cost), you must select "Special Fragmentation" and enter details under the Library Type Specifications field

- RNA low cost high throughput
- Standard RNA Seq Library
- DSN-normalized Library
- Ribosomal Removal
- Poly-A mRNA capture
- Special Fragmentation
- No RNA Enrichment
Library Types

• DNA
  - Standard DNA Library
  - Mate-Pair Library
  - Exome or Custom Capture
  - Special Fragmentation
  - DNA low cost high throughput
  - ddRAD genotyping
  - ddRAD development phase 1
  - ddRAD development phase 2
  - 16S Metagenomics

• RNA
  - RNA low cost high throughput
  - Standard RNA Seq Library
  - DSN-normalized Library
  - Ribosomal Removal
  - Poly-A mRNA capture
  - Special Fragmentation
  - No RNA Enrichment
What Reagents do I use for Library Prep if I want to do it myself?

• There are **many options** to choose from when selecting library prep reagents

• The GSAF mainly uses New England BioLabs (NEB) for all their library prep reagents
  – Manuals are easy to follow
  – Reagents are well priced
  – You can purchase the reagents as a “kit” or as individual modules of the kit, which can be very helpful!

• We’ve also published our protocols in Wiley’s Current Protocols:
  – PMID: 24984855 – DNA
  – PMID: 24733242 - RNA
Library Types & Sequencing

• Important concepts:
  – Single-end vs paired-end sequencing
  – Single-indexing, dual-indexing, and inline indexing

Canonical ILLUUMINA library design as of June 2012 (all 5'-3'), "TruSeq V3": NOTE all sequences shown are TOP STRAND 5' to 3'

Platform descriptions and experimental design suggestions
Dashboard
Big Bio Job Board
Companion Animal Interest Group: Information
How do I start a project with the GSAF?

- Job Submission example
  - Saving vs. Submitting your job

Save and Email – just for you!

Submit Request – goes to GSAF for review/approval
How do I start a project with the GSAF?

• Job Submission example
  – Editing your job BEFORE submitting it
  • You cannot edit your job once you submit it
How can I tell where my project is?

- Your job submission email – link to your job home page

gsaf-requests@utlists.utexas.edu

to jddavis.bme

Your job submission RequestID: 870a53 has been approved and is now Job Number JA14574.

THIS IS AN AUTOMATICALLY GENERATED EMAIL - PLEASE DO NOT REPLY WITH QUESTIONS; contact Jessica Podnar directly if you have any questions.

This is a link to your job home page. It will contain up-to-the-minute information about your job, including how to download your data when it is available.

To avoid a delay or return of your samples to you, please write both your job number and the name of the sample (EXACTLY as entered into the GSAF online submission system) on each tube.

Please note that we have changed our sample drop-off procedures effective March 27, 2013:

- For RNA: Print out this email and drop off this email printout and samples to MBB 3.210 between 11:00 am and 1:00 pm Monday, Wednesday or Friday, or
- For DNA: Print out this email and drop off this email printout and samples to any GSAF staff member MBB 3.210 during normal business hours (8:00am - 5:00pm)
- For questions and concerns, Please send an email to Jessica Podnar or Scott Hunicke-Smith.

To follow your job go to the GSAF Wiki site here, and click this job number in the Job Status Summary table, or click on 'Detailed job information report', and type in your job number.

Thank you,

GSAF
How can I tell where my project is?

• Your job home page

GSAF Data System: Job Home Page for Job JA14227

This page contains up-to-the-minute information about your job. If a section is blank (e.g. QC Information, or Billing information), it means that we have not yet done that task for your job. This information is pulled direction from our production databases, so as soon as the information is available it will appear on this page.

<table>
<thead>
<tr>
<th>Job ID: JA14227</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description: ddRAD protocol development</td>
</tr>
<tr>
<td>Sample receipt date: 17-Mar-14</td>
</tr>
<tr>
<td>Job completion date: 20-May-14</td>
</tr>
</tbody>
</table>

This project has 3 samples and there are 4 QC runs associated with this job.

More information is available below - simply expand the sections to see more.

- QC Information
- Job Summary
- Billing Information
- Sequencing Data (expand to request new data access keys)
- Total sequencing data generated (expand for detail)
How can I tell where my project is?

- The sequencing queues

<table>
<thead>
<tr>
<th>Currently not assigned to a Run</th>
<th>Illumina Barcode</th>
<th>Jake Grohman</th>
<th>JA14743</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Con_p17</td>
<td>=&gt; Illumina Barcode #17</td>
<td>Jake Grohman</td>
<td>JA14743</td>
<td>1</td>
</tr>
<tr>
<td>FHC</td>
<td>=&gt; Illumina Barcode #29</td>
<td>Jake Grohman</td>
<td>JA14743</td>
<td>1</td>
</tr>
<tr>
<td>FHC_DDX5</td>
<td>=&gt; Illumina Barcode #31</td>
<td>Jake Grohman</td>
<td>JA14743</td>
<td>1</td>
</tr>
<tr>
<td>HeLa_p16</td>
<td>=&gt; Illumina Barcode #19</td>
<td>Jake Grohman</td>
<td>JA14743</td>
<td>1</td>
</tr>
<tr>
<td>HT29_p32</td>
<td>=&gt; Illumina barcode #32</td>
<td>Jake Grohman</td>
<td>JA14743</td>
<td>1C</td>
</tr>
<tr>
<td>LeVo_p18</td>
<td>=&gt; Illumina Barcode #16</td>
<td>Jake Grohman</td>
<td>JA14743</td>
<td>1C</td>
</tr>
<tr>
<td>SW480_p28</td>
<td>=&gt; Illumina Barcode #28</td>
<td>Jake Grohman</td>
<td>JA14743</td>
<td>10</td>
</tr>
<tr>
<td>SA14140</td>
<td>3_Smin1frag</td>
<td>TCTTTCCC</td>
<td>GTGAAAAGG</td>
<td>Matthew i</td>
</tr>
<tr>
<td>2x125</td>
<td>Pool1</td>
<td>multiple</td>
<td>multiple</td>
<td>Matthew i</td>
</tr>
<tr>
<td>Start Date:</td>
<td>22_06_031</td>
<td>TAATCTTA</td>
<td>TCCGCGAA</td>
<td>Matthew i</td>
</tr>
<tr>
<td>slide ID:</td>
<td>22_06_015</td>
<td>TAATCTTA</td>
<td>TAATCGGC</td>
<td>Matthew i</td>
</tr>
<tr>
<td>Sequencer:</td>
<td>22_06_021</td>
<td>ATAGAGGC</td>
<td>CGCGCTATG</td>
<td>Matthew i</td>
</tr>
<tr>
<td>Barcodes:</td>
<td>22_06_021</td>
<td>GGCCTCGA</td>
<td>CGCGCTATG</td>
<td>Matthew i</td>
</tr>
<tr>
<td>DUAL INDEX</td>
<td>22_06_043</td>
<td>TATAGCTG</td>
<td>AGCGATAAG</td>
<td>Matthew i</td>
</tr>
<tr>
<td></td>
<td>r3_1</td>
<td>ATACGC</td>
<td>LIN QIU</td>
<td>JA14637</td>
</tr>
<tr>
<td></td>
<td>r2_11</td>
<td>TGACCA</td>
<td>LIN QIU</td>
<td>JA14637</td>
</tr>
<tr>
<td></td>
<td>r3_11</td>
<td>TTAGGC</td>
<td>LIN QIU</td>
<td>JA14637</td>
</tr>
</tbody>
</table>
How do I get my data?

- Your data delivery email comes to you automatically

GSAF Data Key Available

New access keys have been generated for each of your data files. Anyone with access to your data keys will be able to download your data. We have limited the time window for download to 72 hours.

Access your data for JA14699 from sequencing run SA14130 here.

This link will be accessible for ten days from the date of this email.

You may request new access keys and find many details about your job on your job home page.

Please see this page for more information about downloading your data.

Sincerely,

GSAF

This email was auto-generated by data delivery system version 0.0.2 (alpha) of the UT GSAF.
How do I get my data?

- Your job home page – you can access ANYTIME

GSAF Data System: Job Home Page for Job JA14699

This page contains up-to-the-minute information about your job. If a section is blank (e.g., QC Information or Billing information), it means that we have not yet done that task for your job. This information is pulled directly from our production database.

<table>
<thead>
<tr>
<th>Job ID: JA14699</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description: heavy chain amplicons (~400 bp) to be sequenced by 2 x 300 bp MiSeq</td>
</tr>
<tr>
<td>Sample receipt date: 02-Sep-14</td>
</tr>
<tr>
<td>The job has not been completed</td>
</tr>
<tr>
<td>This project has 23 samples and there are 4 QC runs associated with this job.</td>
</tr>
<tr>
<td>More information is available below - simply expand the sections to see more.</td>
</tr>
</tbody>
</table>

**QC Information**

- **Job Summary**
- **Billing Information**
- **Sequencing Data (expand to request new data access keys)**

Sequencing run SA14130: (select this link to generate access keys for SA14130)

<table>
<thead>
<tr>
<th>File</th>
<th># of read-pairs</th>
<th>md5sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>3173poMIM_S1_L001_L1_001.fastq.gz</td>
<td>818911</td>
<td>ee328eaff02e5c9a1ec4dcd81a8d41d7</td>
</tr>
<tr>
<td>3173poMIM_S1_L001_R1_001.fastq.gz</td>
<td>818911</td>
<td>1d5b7ad4823722d486ecb3f27ccba411</td>
</tr>
<tr>
<td>3173poMIM_S1_L001_R2_001.fastq.gz</td>
<td>818911</td>
<td>3353e10330d5ecd45bc5b7fa39993911</td>
</tr>
<tr>
<td>3173poMock_S2_L001_L1_001.fastq.gz</td>
<td>694653</td>
<td>08b9c1da413b7d0523a0aafa3f5c9734</td>
</tr>
<tr>
<td>3173poMock_S2_L001_R1_001.fastq.gz</td>
<td>694653</td>
<td>5d2f88d02b3942c2344a61b2e1982e3</td>
</tr>
<tr>
<td>3173poMock_S2_L001_R2_001.fastq.gz</td>
<td>694653</td>
<td>037458b8b3a6c5b8d481e5d070d32</td>
</tr>
</tbody>
</table>

Clicking this link will re-generate an email with new data keys to any email addresses listed on the project.
Frequently Asked Questions
Example Projects

• RNA-seq, poly-A mRNA enrichment for differential expression
• Bacterial whole-genome assembly
• ChIP-seq
• ddRAD
Thank you from the GSAF!

The GSAF: Jessica Podnar, Matt Barnette, Heather Diedrick, Terry Heckman, Gabby Huerta, Mani Singh