Sequencing @ Berkeley

October 3, 2012

10am-12pm

238 Koshland Hall

Genomics Facilities Introduction

Functional Genomics Laboratory

Vincent J. Coates Genomic Sequencing Facility

Computational Genomics Resource Laboratory

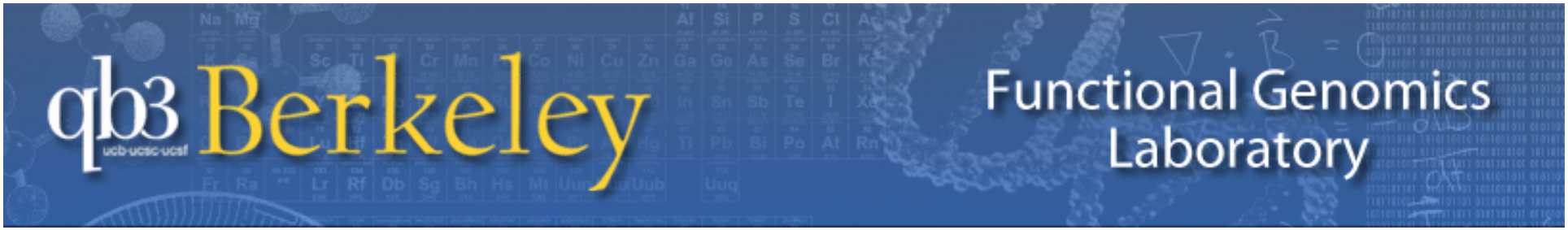
Functional Genomics Laboratory (FGL)

The FGL is a core facility of the California Institute for Quantitative Biosciences and is open to both university and industrial users.

The Functional Genomics Laboratory enables researchers to conduct state-of-the-art research in functional genomics.

Complementing the high-throughput DNA sequencing capabilities of the GSL, the FGL specializes in the quality control of the starting & final product of library prep. We are also actively involved in the development of new techniques for NGS experimentation.

Research support services include hands-on training and consultation.



Vincent J. Coates
Genomic Sequencing Facility (GSL)

- Located in B206 Stanley Hall
- Currently supporting the Illumina Platform
  - 3 HiSeq2000s
  - 3 cBot (cluster generation)
- Staff
  - Minyong Chung
  - Karen Lund (50% time)

Faculty Advisor

- Don Rio

qb3 ucb-ucsc-ucsf Berkeley V.J.C. Genomics Sequencing Lab

Computational Genomics Resource Laboratory

- Core facility within QB3 institute – 2010
- Computational resources for analysis of next generation sequence data
- Faculty directors
  - Brian Staskawicz, John Taylor
- Staff
  - Madhavan Ganesh
  - Donna Hendrix (part-time)

What is Next Gen Sequencing?

- Highly parallelized sequencing
- Ability to sequence full genomes in one run
- Low hands-on time, but huge volumes of data
- Highly Accurate
- Many different platforms available
  - Illumina, SMRT/PacBio,
    Ion Torrent+Solid/Life, 454/Roche

Next Generation Seq Experiments

Experiment Design & Library Preparation

DeNovo
Whole Genome
Targeted Reseq.
RNA-Seq
Chip-Seq
BS-Seq
RAD-Seq, Etc.

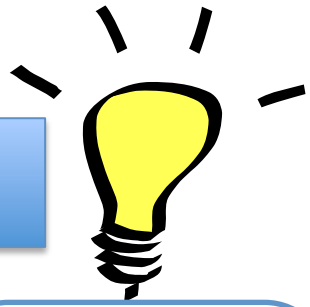
Sequencing

Sequence Data

Fastq format data files (Fasta with quality scores)

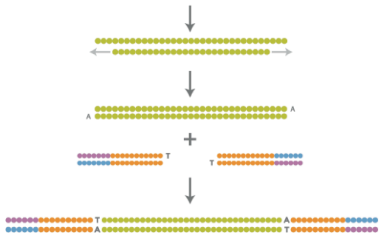
Data Analysis

Insights



Novel Discoveries→
Publications→
Prestigious Awards→
Save/Change the world

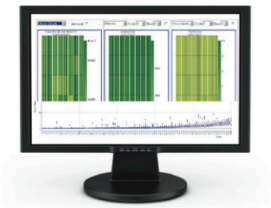
Illumina Platform Workflow



Library Preparation
~2 h [15 min hands-on (Nextera)]
< 6 h [< 3 h hands-on (TruSeq)]

Cluster Generation
~5 h (<10 min hands-on)

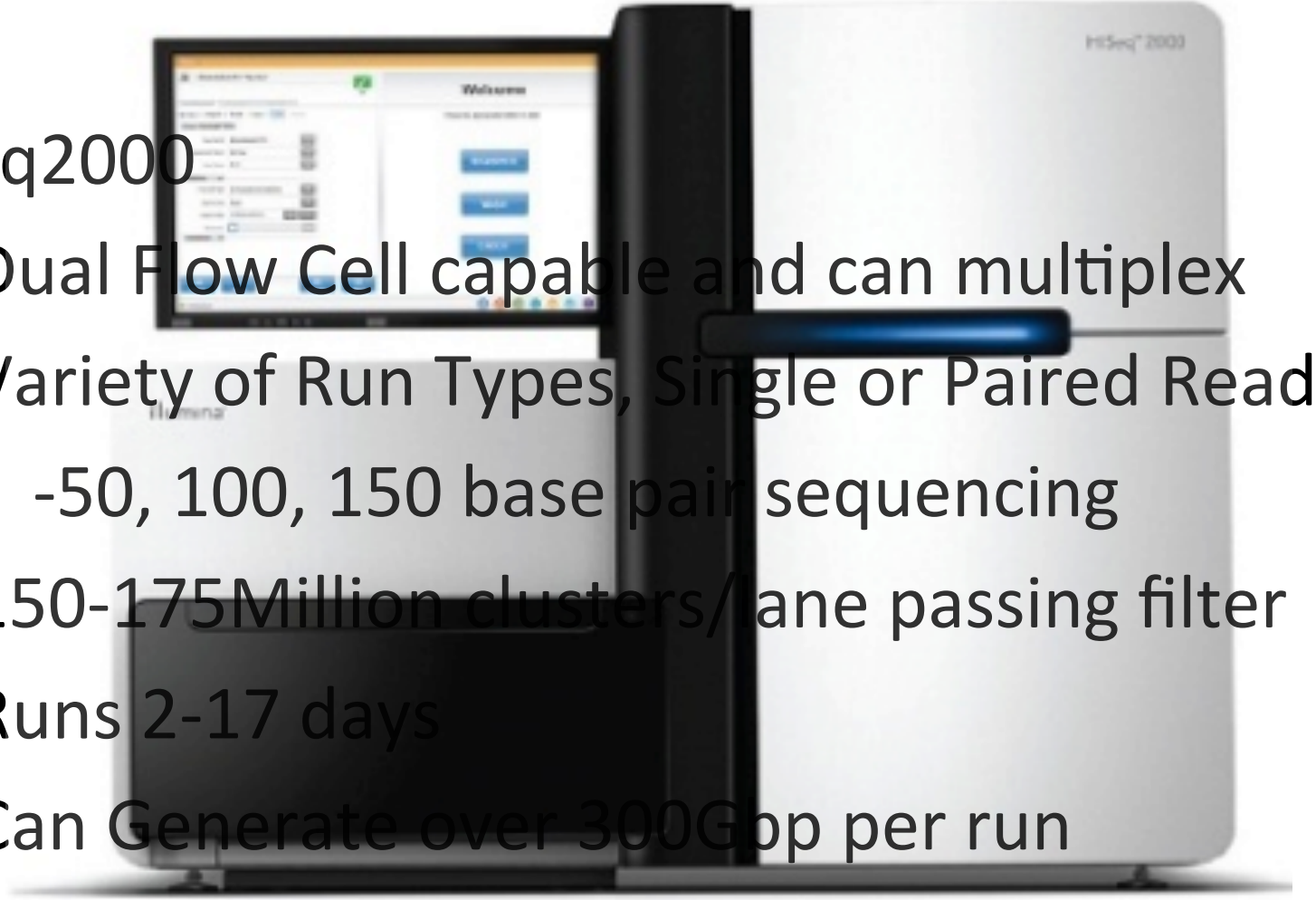
Sequencing by Synthesis
~1.5 to 11 days



CASAVA
2 days (30 min hands-on)

- Sample Library Preparation → FGL
- Sample Sequencing → GSL
- Data Analysis → CGRL

Illumina Platform Statistics



HiSeq2000

-Dual Flow Cell capable and can multiplex

-Variety of Run Types, Single or Paired Reads

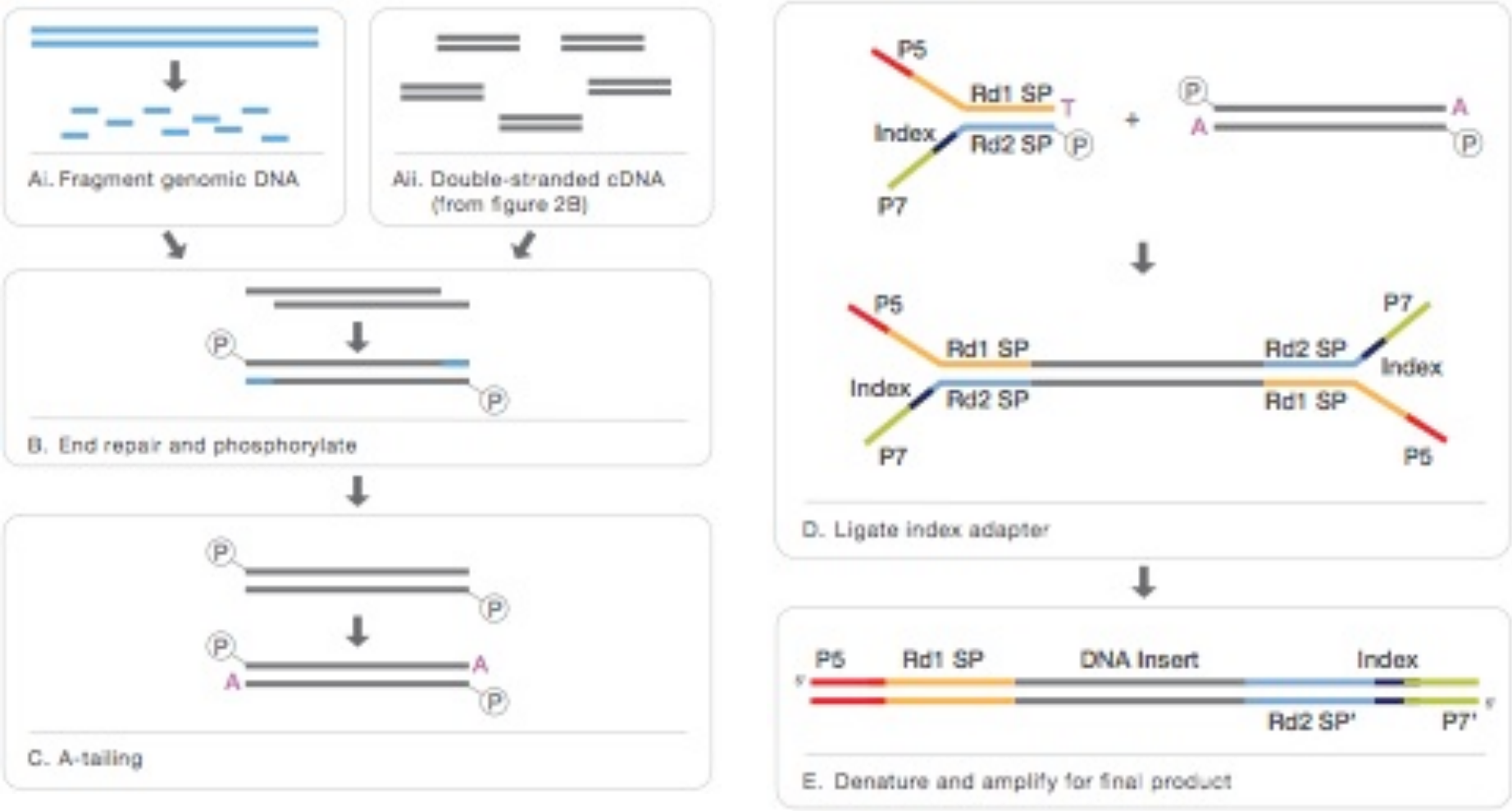
-50, 100, 150 base pair sequencing

-150-175Million clusters/lane passing filter

-Runs 2-17 days

-Can Generate over 300Gbp per run

Sample Library Preparation



Sample preparation Kits

- Illumina's Truseq and Nextera
- Bioo Scientific's Air/NEXTflex
- NuGen's Encore/Ovation
- NEB's NEBNext
- Kapa Biosystem's Library prep kits
- Epicentre's ScripSeq (RNA)
- Many, many others

GSL's Sample Sequencing workflow

Submit Illumina Libraries

Sample QC

Cluster Sample

Sequencing

Data Delivery

Receive Illumina Libraries

B206 Stanley

255 LSA

Qubit

Bioanalyzer

qPCR

Clonal Amplification of libraries onto Flow Cell (4-5 hrs)

Run on HiSeq200 (3-17 days)

Process Data and FTP Delivery (1 day)

qb3 ucb-ucsc-ucsf Berkeley V.J.C. Genomics Sequencing Lab

Cluster Amplification

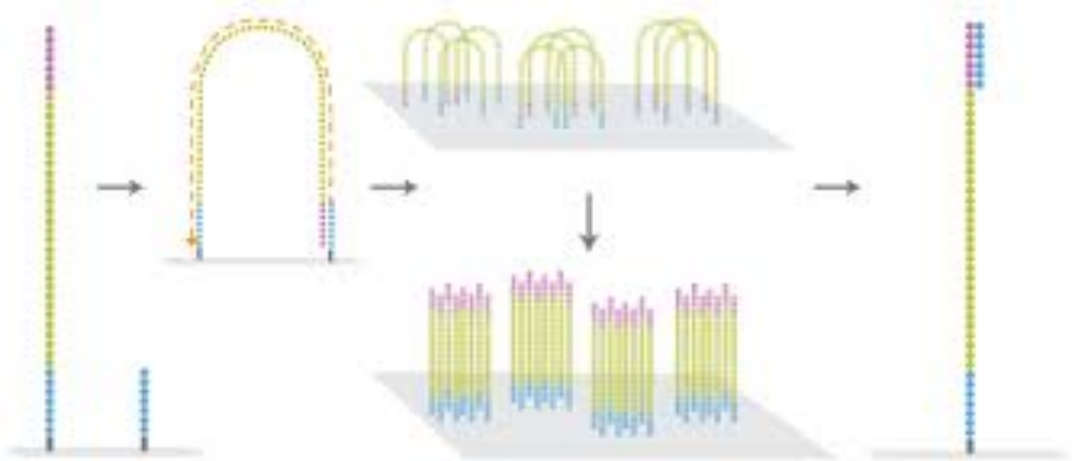
Cluster Sample

Clonal Amplification of libraries onto Flow Cell (4-5 hrs)

cBot –Clonal Amplification Robot

→ SR or PE flow cells

→ Samples denatured to pM range and set in cBot for clonal amplification



Sequencing and Data Delivery

Sequencing

Run on HiSeq200 (3-17 days)

HiSeq2000

-Sequence by Synthesis

→Fluorescently labeled reversible terminator bases

-Need full 8 lanes to run

-Data is generated by RTA to minimize processing post-run

Data Delivery

Process Data and FTP Delivery (1 day)

Post Run

-Basecall files converted to FASTQ

-Run QC to ensure Sequencing ran without machine failures.

-Data sets QC'ed to Illumina's specs.

-Delivery via FTP

-Notification via email.

GSL Recharge fees

HiSeq2000 Sequencing costs per lane (UC):

	Single-End Reads	Paired-End Reads
50bp	$873	$1526
100bp	$1268	$1869
150bp	$1607*	$2547

→Current Turnaround times 4-7 weeks (run specific)

→ GSL Averages 15-17 runs/month

→ GSL Receives about 400 sample tubes/month

Future Plans

4th HiSeq2000

→ Increase capacity to shorten turnaround time

HiSeq2500 Upgrade:

→Rapid Run mode: 2 lanes of 150PE in 3 days

- only 80% of the data of a HiSeq2000 lane
- approximately 20% more expensive
- FAST

Expansion of Servers

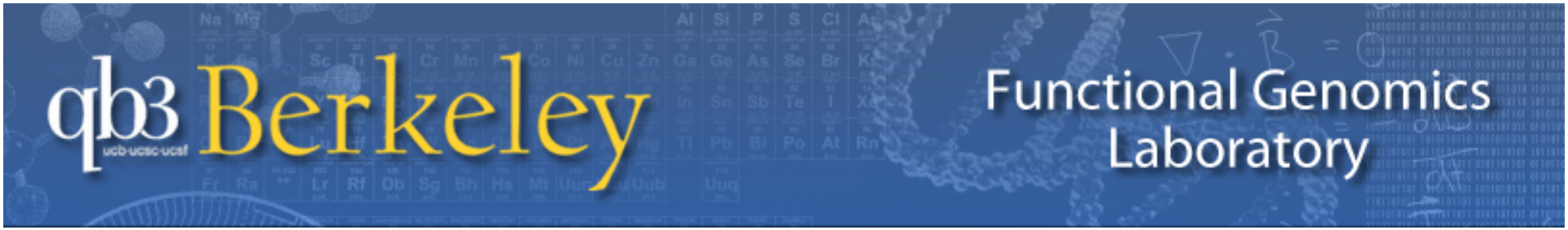
Functional Genomics Laboratory (FGL)

The FGL is a core facility of the California Institute for Quantitative Biosciences and is open to both university and industrial users.

The Functional Genomics Laboratory enables researchers to conduct state-of-the-art research in functional genomics.

Complementing the high-throughput DNA sequencing capabilities of the GSL, the FGL specializes in the quality control of the starting & final product of library prep. We are also actively involved in the development of new techniques for NGS experimentation.

Research support services include hands-on training and consultation.



Available services for NGS at FGL

1. Qubit
2. Bioanalyzer
3. Covaris
4. Library prep

Qubit® Fluorometer

- More accurate quantitation via strand-specific dye absorption

Qubit® Assay Kit	Assay Range	Sample Starting Concentration
dsDNA HS Assay	0.2–100 ng	10 pg/µL–100 ng/µL
dsDNA BR Assay	2–1000 ng	100 pg/µL–1 ug/µL
RNA Assay	5–100 ng	250 pg/µL – 100 ng/µL

Covaris: shearing by sonication

Uses soundwaves

- No sample contamination with buffers or enzymes
- Mechanical shearing at random, more even coverage
- Not concentration-specific

Uses:

- Fragment DNA evenly for increased library complexity
- ChIP

Bioanalyzer

Sample composition analysis

- •Fragment sizes
- •Concentration and molarity of separate products within sample

Analysis of sample Integrity

- •RIN score of RNA samples
- •Morphology of peaks

Useful at many different stages of experimental workflow

- •Evaluate quality of starting material
- •Check efficacy of fragmentation
- •Composition of final product, efficacy of cleanup

Reproducible, quick, not messy, wealth of analytical output

Current available Assays

DNA Assays:

DNA1000 DNA7500 DNA HS

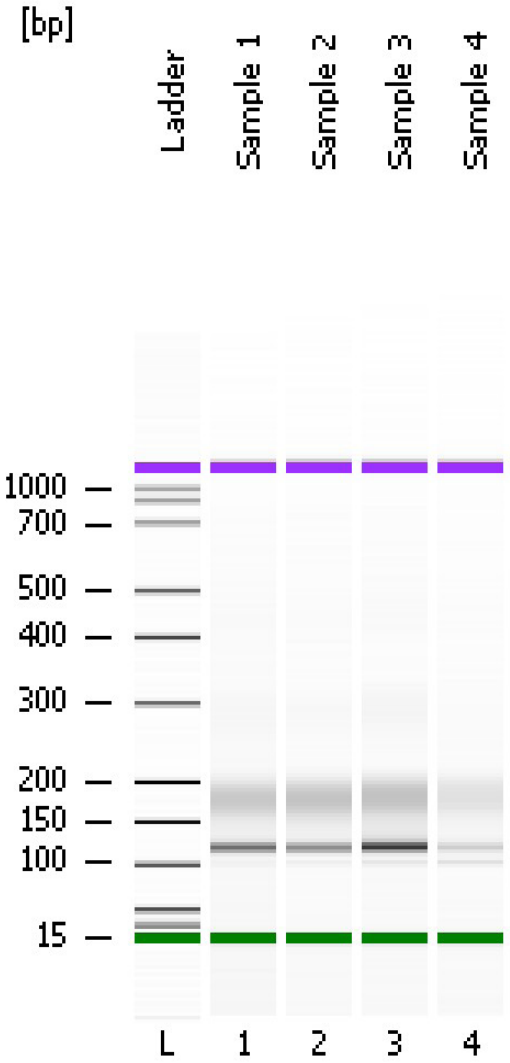
- Sizing
- Quantitation
- PCR products, digests, larger DNA fragments
- 11-12 samples in 45 min.

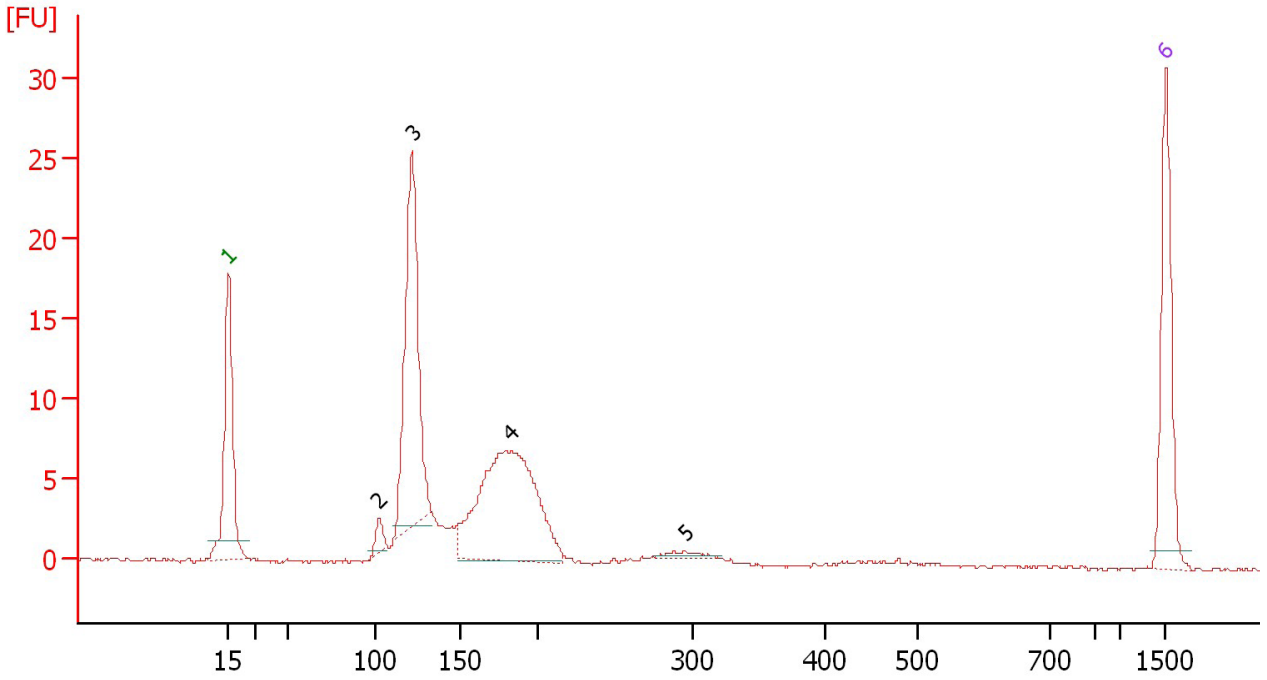
RNA Assays:

6000 Nano 6000 Pico Small RNA

- Quantitation (Sizing in Small RNA)
- Total RNA, mRNA
- Purity & integrity determination
- 11-12 samples in 30 min.

DNA Assays





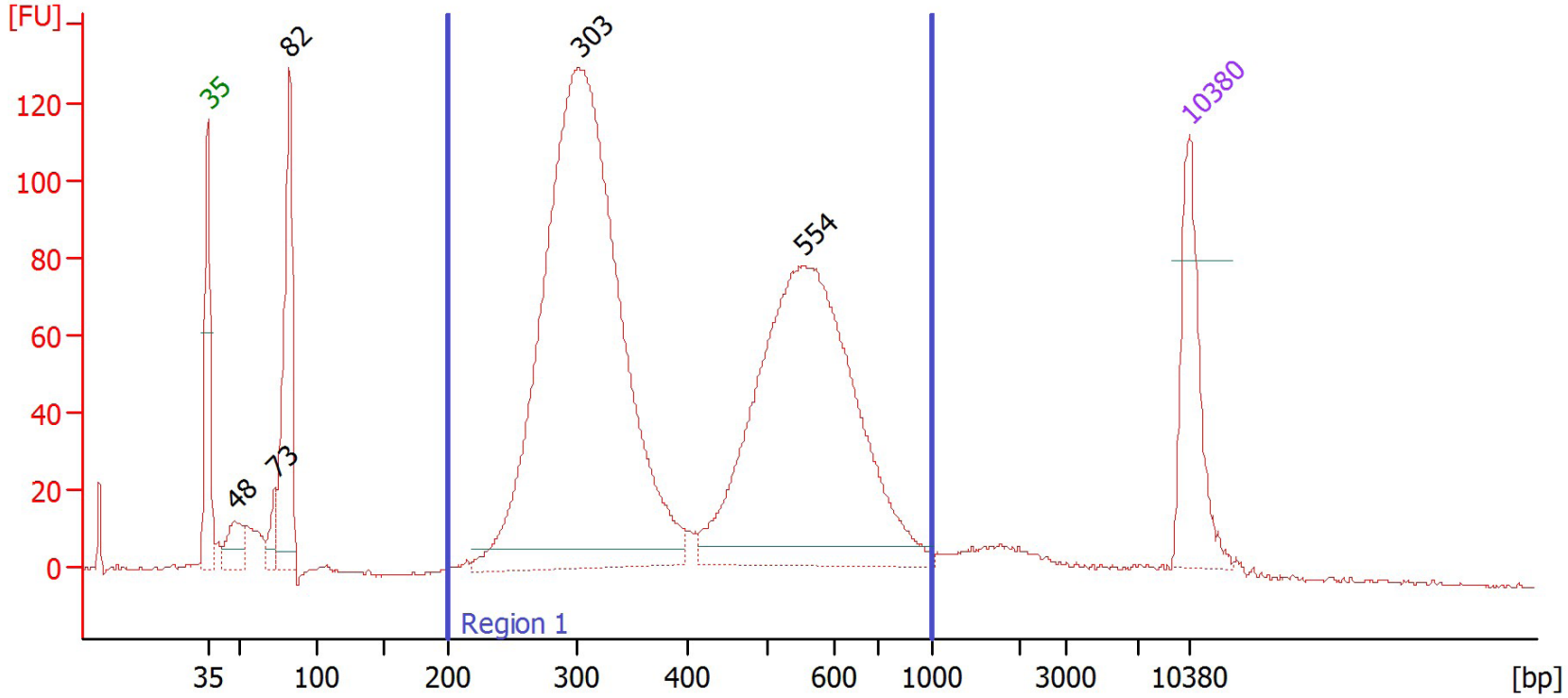
Overall Results for sample 3 : **Sample 3**

Peak table for sample 3 : **Sample 3**

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]
1	15	4.20	424.2
2	102	0.26	3.9
3	121	4.74	59.4
4	182	5.40	44.9
5	294	0.17	0.9
6	1,500	2.10	2.1

Number of pea

DNA Assays

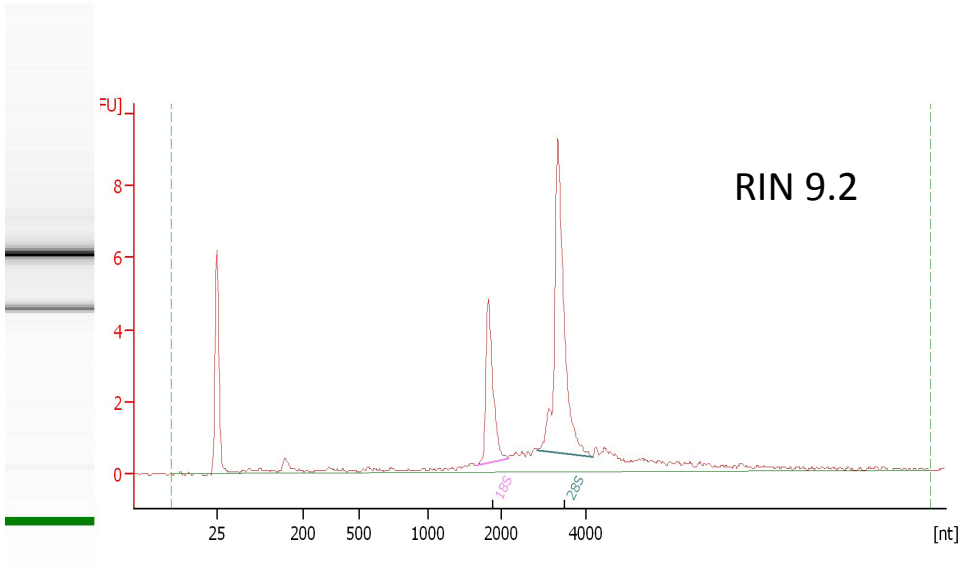


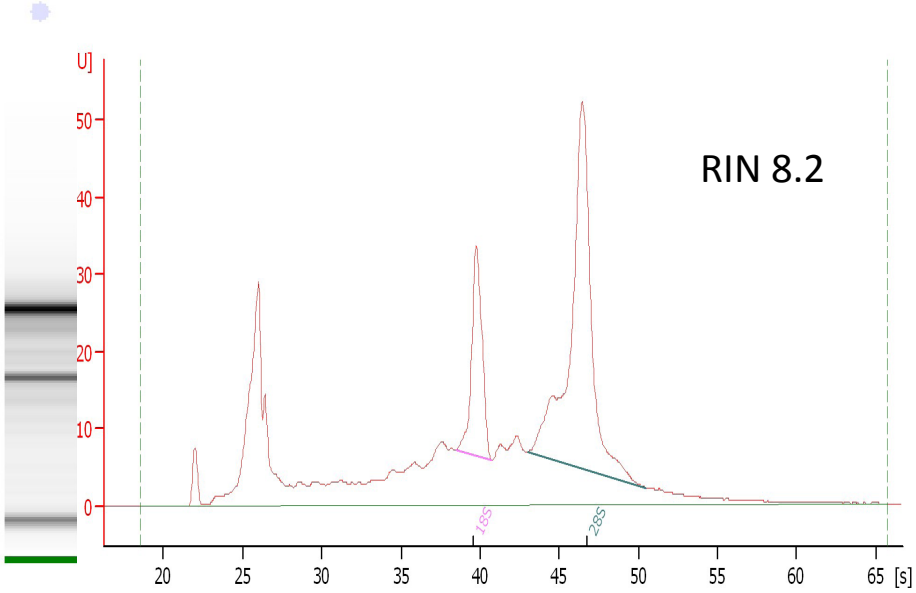
Region table for sample 7 : 4+14

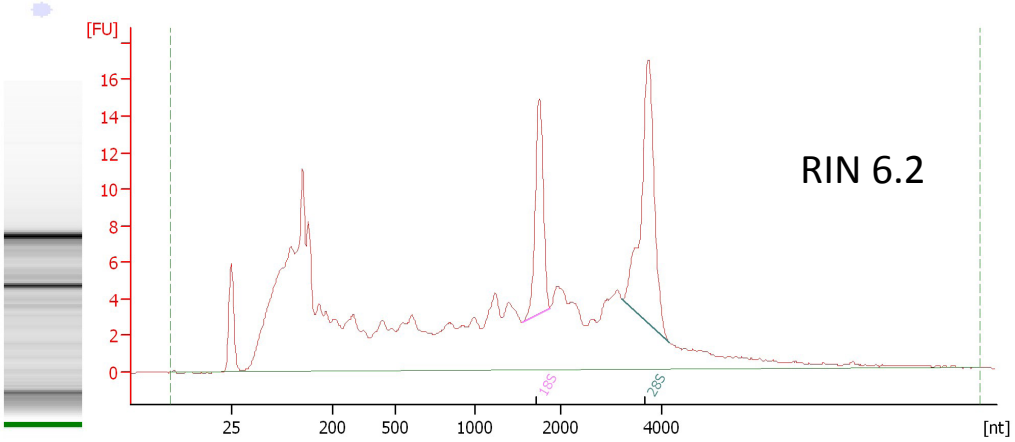
From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]
200	1,000	2,145.8	83	427	34.9	1,551.32	6,470.0

Total RNA Pico Assay

Electropherogram

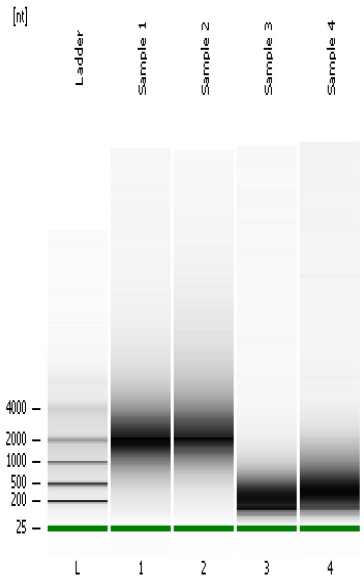


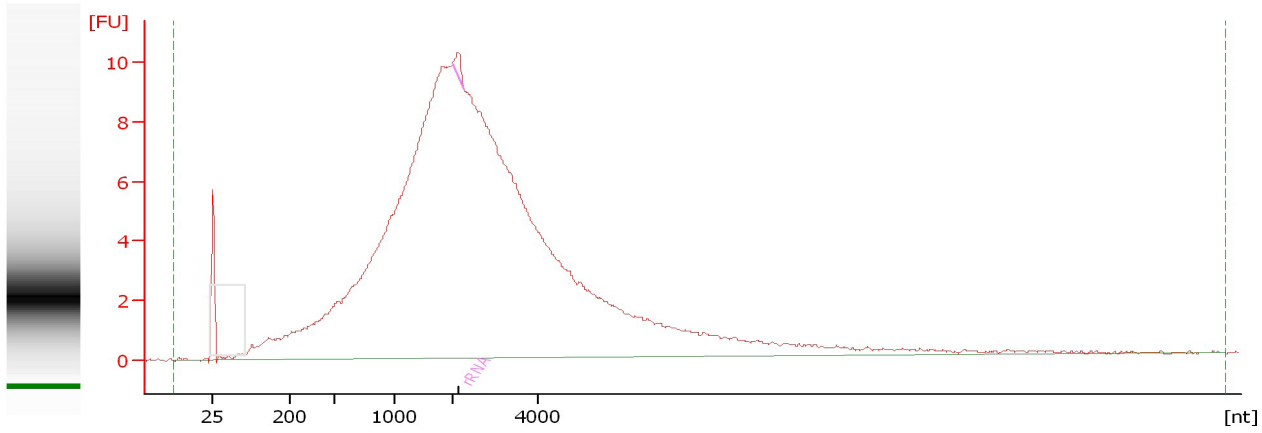




RNA Area: 439.6
RNA Concentration: 7,938 pg/µl
rRNA Ratio [28s / 18s]: 1.9
RNA Integrity Number (RIN): 9.2

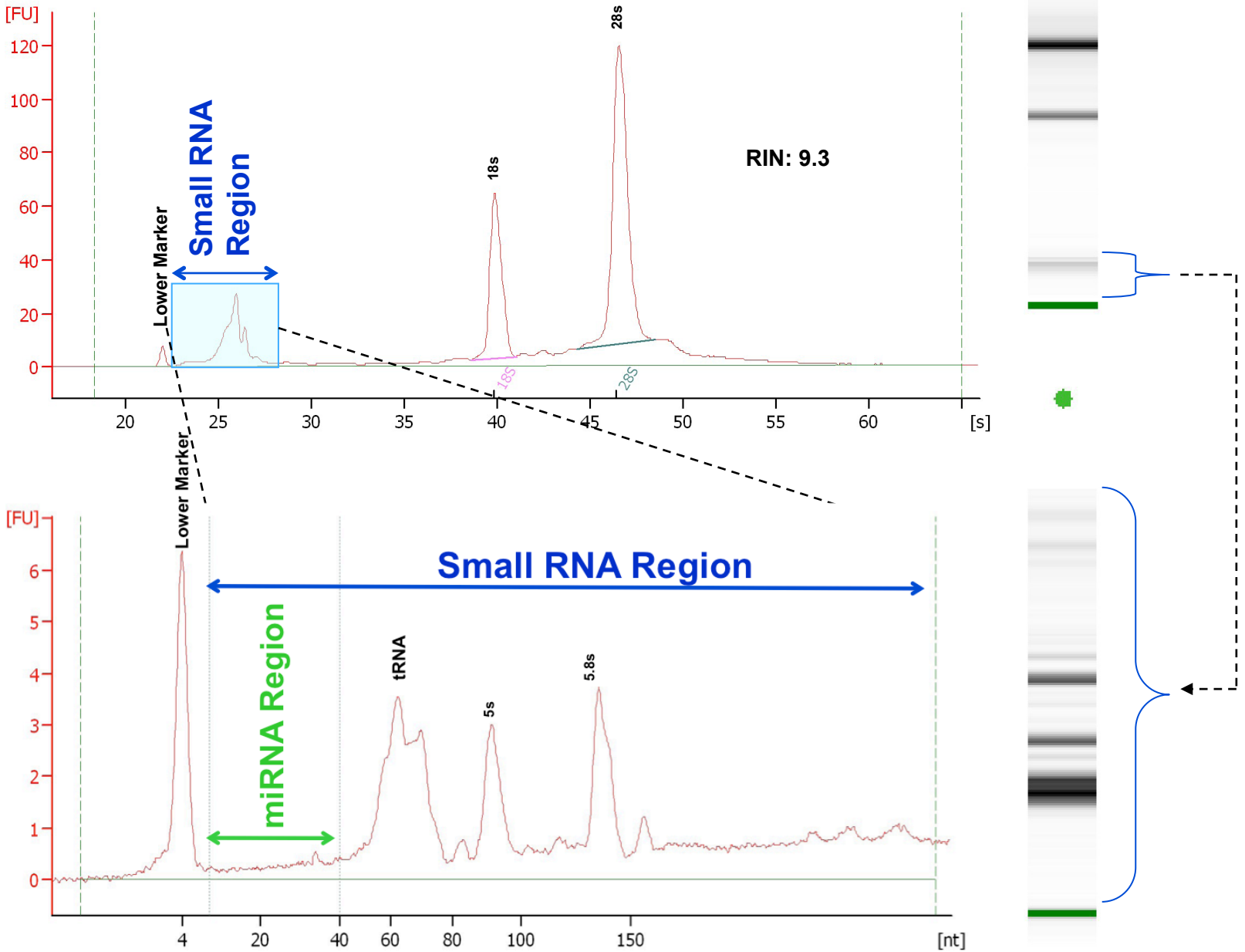
mRNA Pico assay



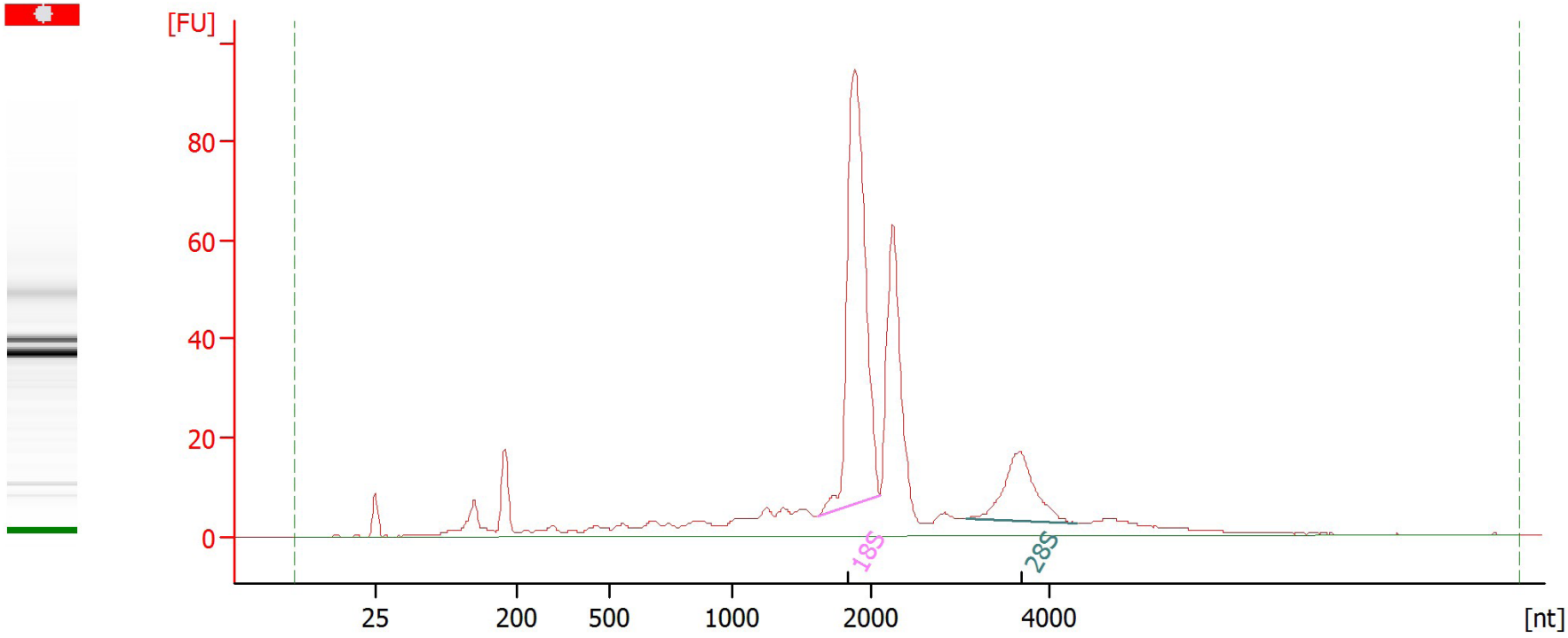


RNA Area: 320.1
RNA Concentration: 3,230 pg/µl
rRNA Contamination: 0.3 %

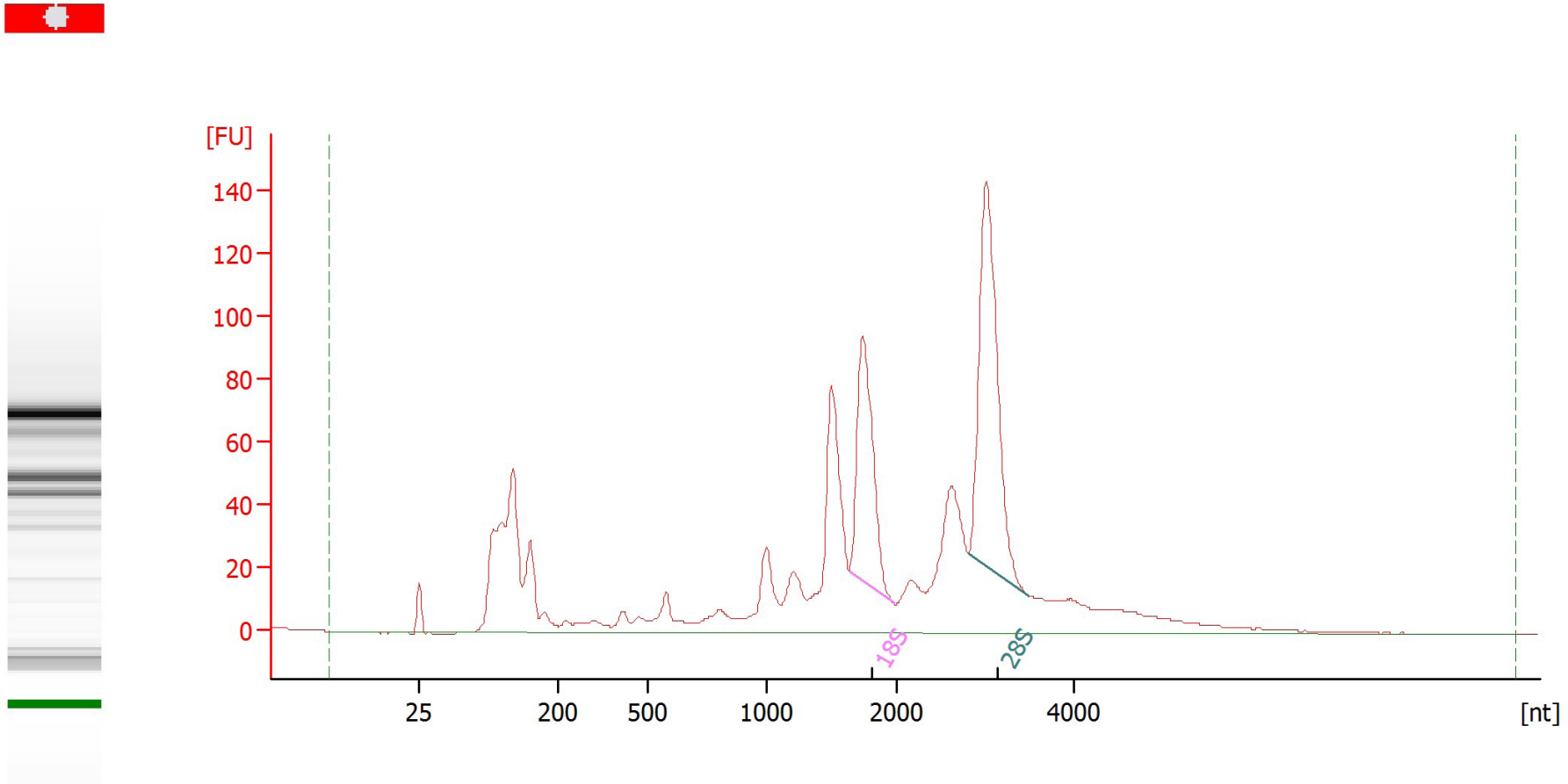
smRNA Assay


[FU]
120
100
80
60
40
20
0
Small RNA Region
Lower Marker
18s
28s
RIN: 9.3
20
25
30
35
40
45
50
55
60
[s]
[FU]
6
5
4
3
2
1
0
Lower Marker
Small RNA Region
miRNA Region
tRNA
5s
5.8s
4
20
40
60
80
100
150
[nt]

Common Fly Profile

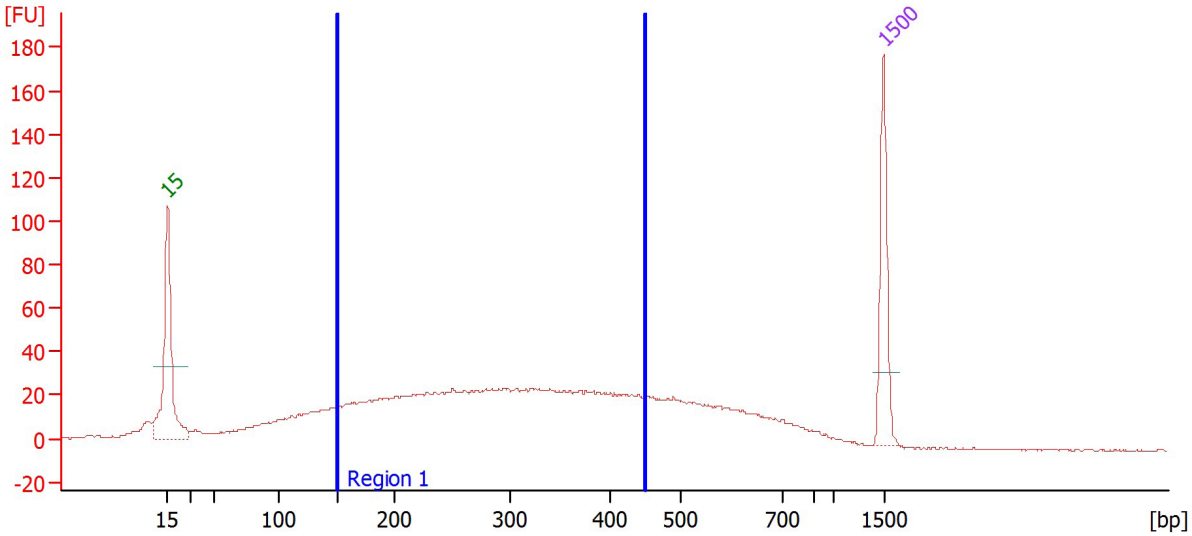


Plant Profile



Using Bioanalyzer to evaluate sheared starting material

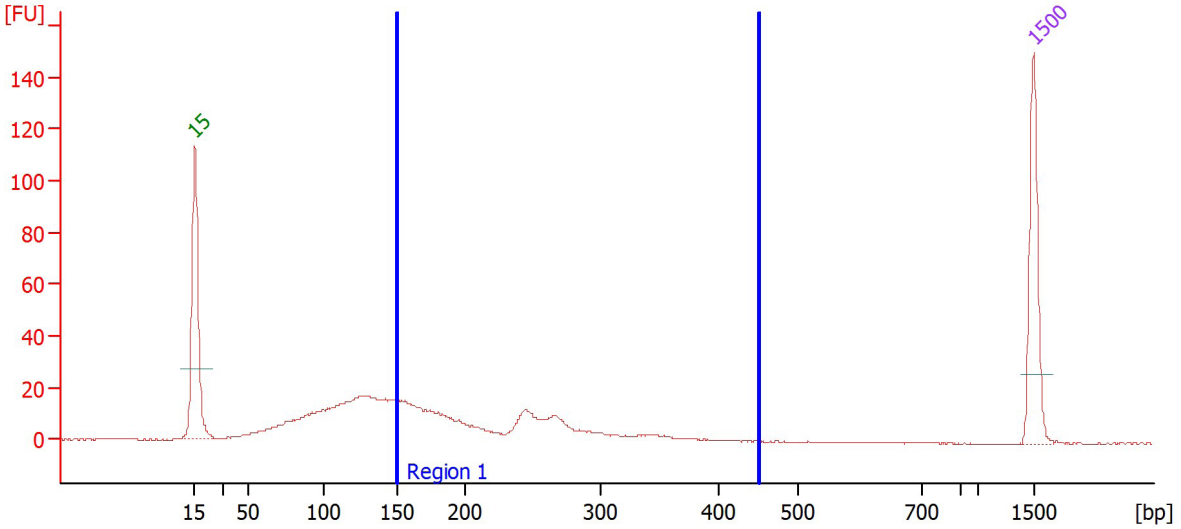
Good:
• evenly sheared within sizing range



Bad:

•Incomplete shearing within sizing range

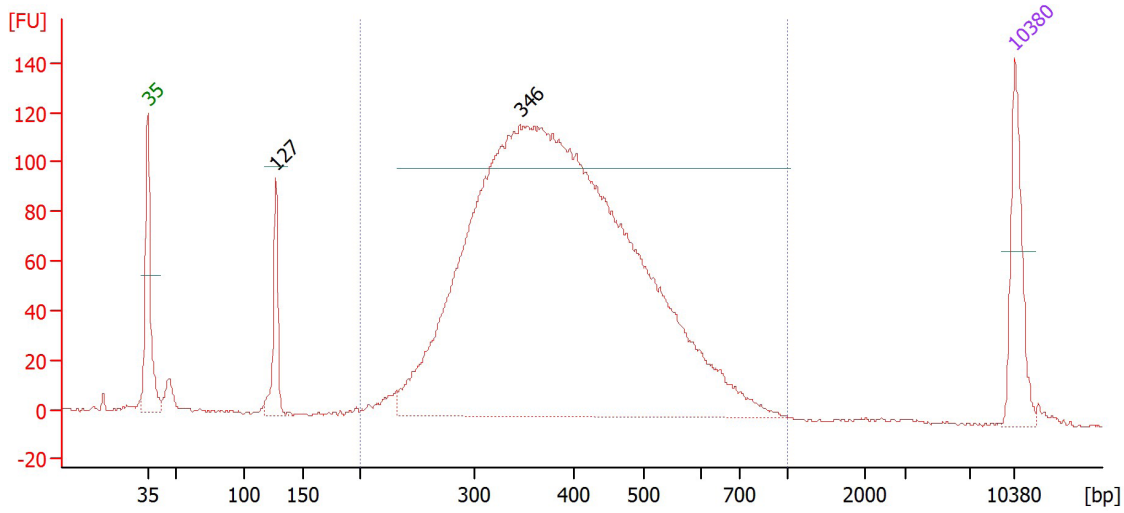
•Much of product too small



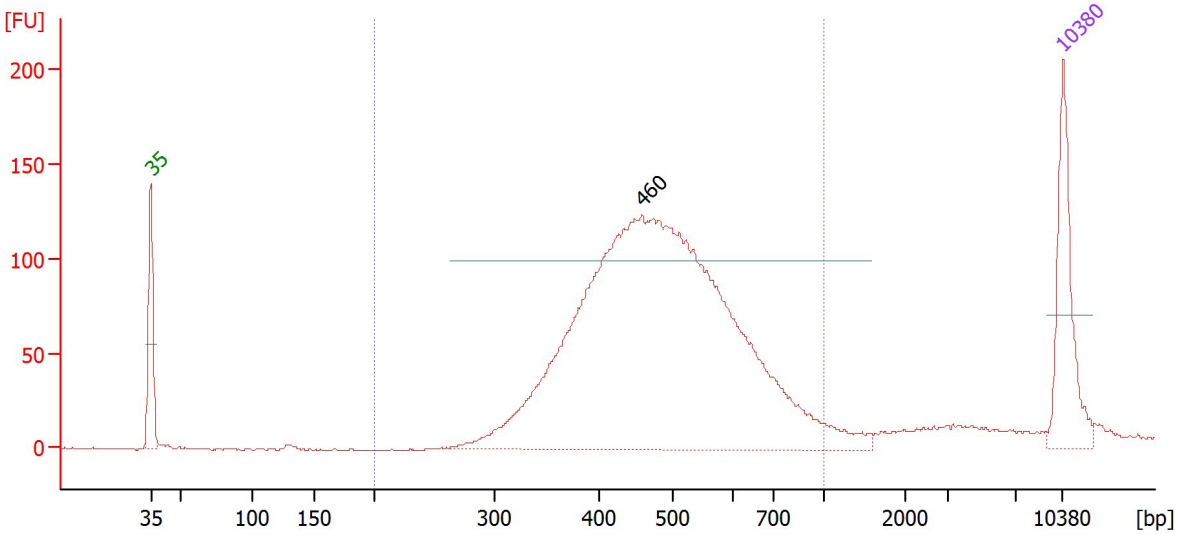
Evaluation of library clean-up

Size [bp]	Conc. [pg/µl]	Molarity [pmol/l]
35	125.00	5,411.3
127	90.34	1,074.9
346	2,160.32	9,460.6
10,380	75.00	10.9

Adapter-to-library ratio:
~9.5% adapters



Good cleanup:
no peaks



Library Prep

Workflow	Function	Services included
dsDNA	Submission of sample template	
Fragmentation		Covaris, Bioanalyzer
Apollo	Size selection, adapter ligation	Qubit
PCR	Increase of ligated product proportion, Indexing if needed	
Cleanup	Sequencing-ready library	Bioanalyzer

About Apollo system

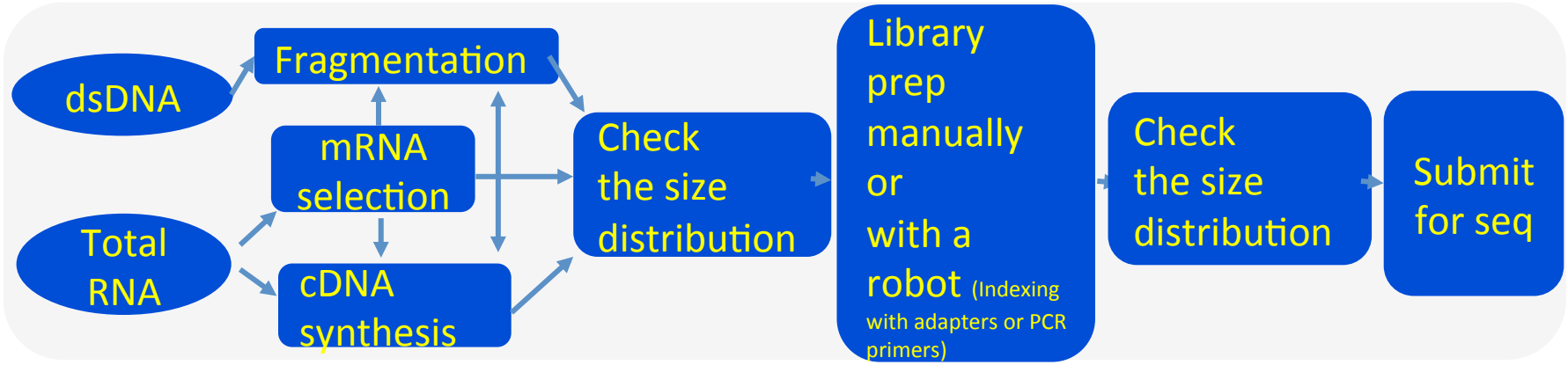
Increases consistency for projects with many samples

Includes everything needed for size selection and adapter ligation steps

PrepX™ Library Kit	Sequencing Platform	Input	Fragmentation	Input Amount	Input Volume	# samples per run	Run time (hours)	Output Size	Catalogue Number
PrepX DNA	Illumina® GA, HiSeq™, or MiSeq™	Fragmented DNA	Covaris® or Nebulizer	1 ng – 3 µg*	15 µL	1 to 8	3.5	255 – 405 bp	400007
PrepX 32 DNA	Illumina GA, HiSeq, or MiSeq	Fragmented DNA	Covaris or Nebulizer	1 ng – 3 µg*	15 µL	9 to 32	7.5	255 – 405 bp	400021
PrepX PGM DNA	Ion Torrent™ PGM™	Fragmented DNA	Covaris or Nebulizer	300 ng – 5 µg*	15 µL	1 to 8	3.5	150 – 250 bp	400029
PrepX 454 DNA	Roche 454 GS FLX Titanium™ or GS Junior™	Fragmented DNA	Nebulizer	500 ng	15 µL	1 to 8	1.5	350 – 1000 bp	400008
	Roche 454 GS FLX+	Fragmented DNA	Nebulizer	500 ng	15 µL	1 to 8	1.5	650 – 3000 bp	
PrepX BeadX High (for ChIP sequencing)	Illumina GA, HiSeq, or MiSeq	Fragmented DNA	Covaris or Nebulizer	1 ng – 3 µg*	15 µL	1 to 8	3.5	> 200 bp	400024
PrepX SPIA® RNA Seq	All	Total RNA	Library specific	2 – 10 ng	5 µL	1 to 8	5.5	N/A (cDNA)	400030

**Field tested inputs.*

Overview - FGL Instruments



Available Instruments	Covaris S220 Apollo	Bioanalyzer	Apollo PCR	Bioanalyzer	HiSeq 2000
Available Services	Covaris RNA-Seq library prep	Bioanalyzer	DNA Library prep Robot	Bioanalyzer	Sequencing

qb3 Berkeley Functional Genomics Laboratory

Future Upcoming Services

Suggestions?

Possible future directions if users are interested:

- Directional RNA-seq
- MicroRNA-seq
- Pippin-Prep size selection service
- Bisulfite library prep

qb3 Berkeley

Computational Genomics Resource Laboratory

NGS Data Analysis - CGRL

Madhavan Ganesh

Oct 03, 2012

CGRL Fall 2012 Workshops

Outline of Talk

- CGRL's role in the analysis of NGS data
- Resources available for data analysis at CGRL
- Fall Workshops – upcoming workshops
- Q & A / Contact Information

CGRL Mission

- Train biologists to do genomic data analysis
  - Short courses – computational tools
    - Unix, Perl, Python, R
  - Workshops, Seminars – analysis methods
    - RNASeq, Assembly, ChIPSeq, SNP/Genotyping analysis
  - One-on-one help as resources permit
- Provide computational resources for data analysis
  - Compute cluster
  - Storage (not long term!)
  - Software

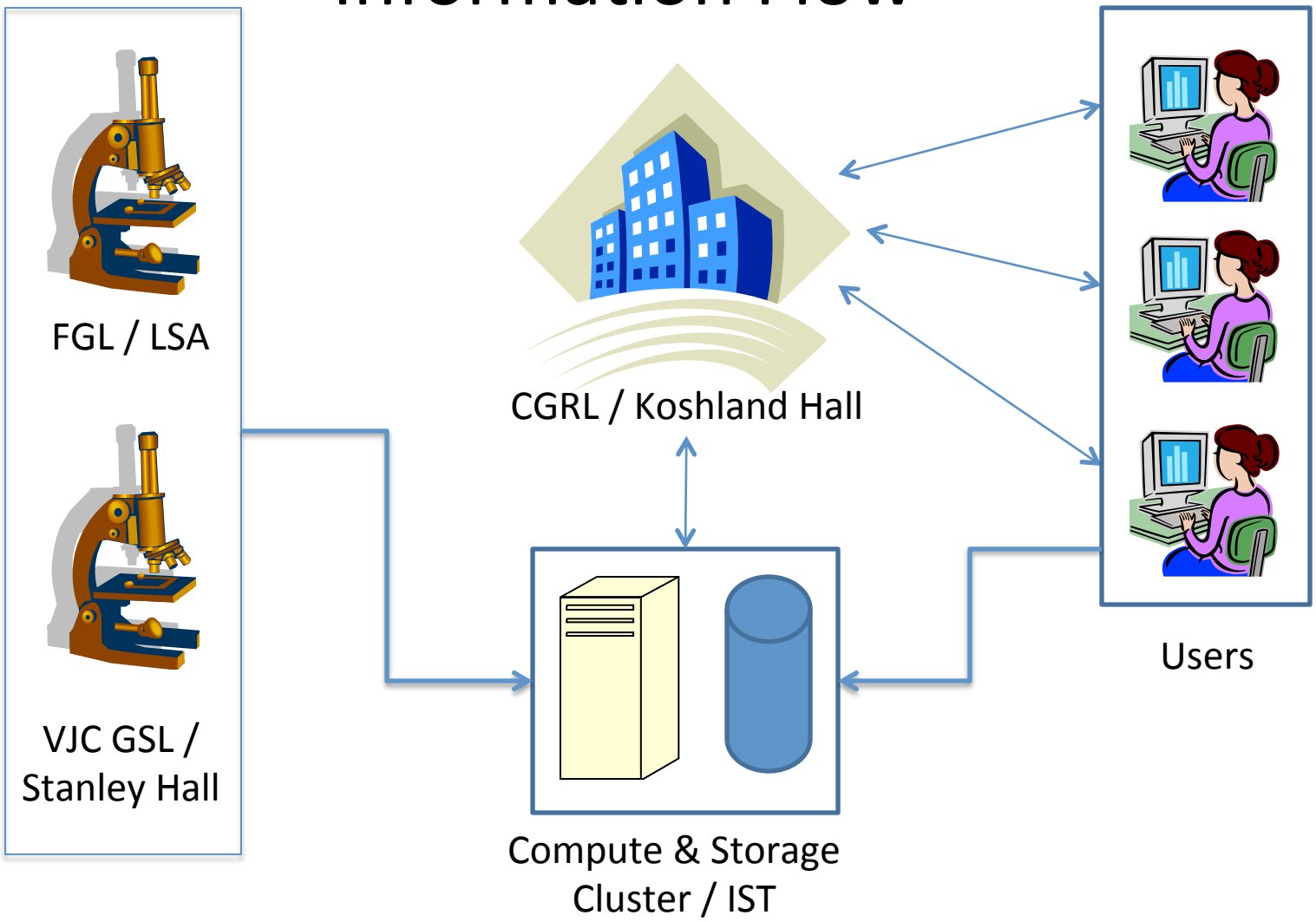
NGS Experiments: Prepare for Analysis

- Give due emphasis to
  - Experiment Design
    - Statistical validation, replicates, sequencing depth
  - Data Processing & Data Management
- Data Processing
  - Volume of sequence data output
    - few 10s of GB per lane from Illumina HiSeq
  - Desktop office tools are not sufficient
    - Word, Excel etc.
  - Some expertise with Unix environment
    - Scripting (Shell, Perl/Python, R)

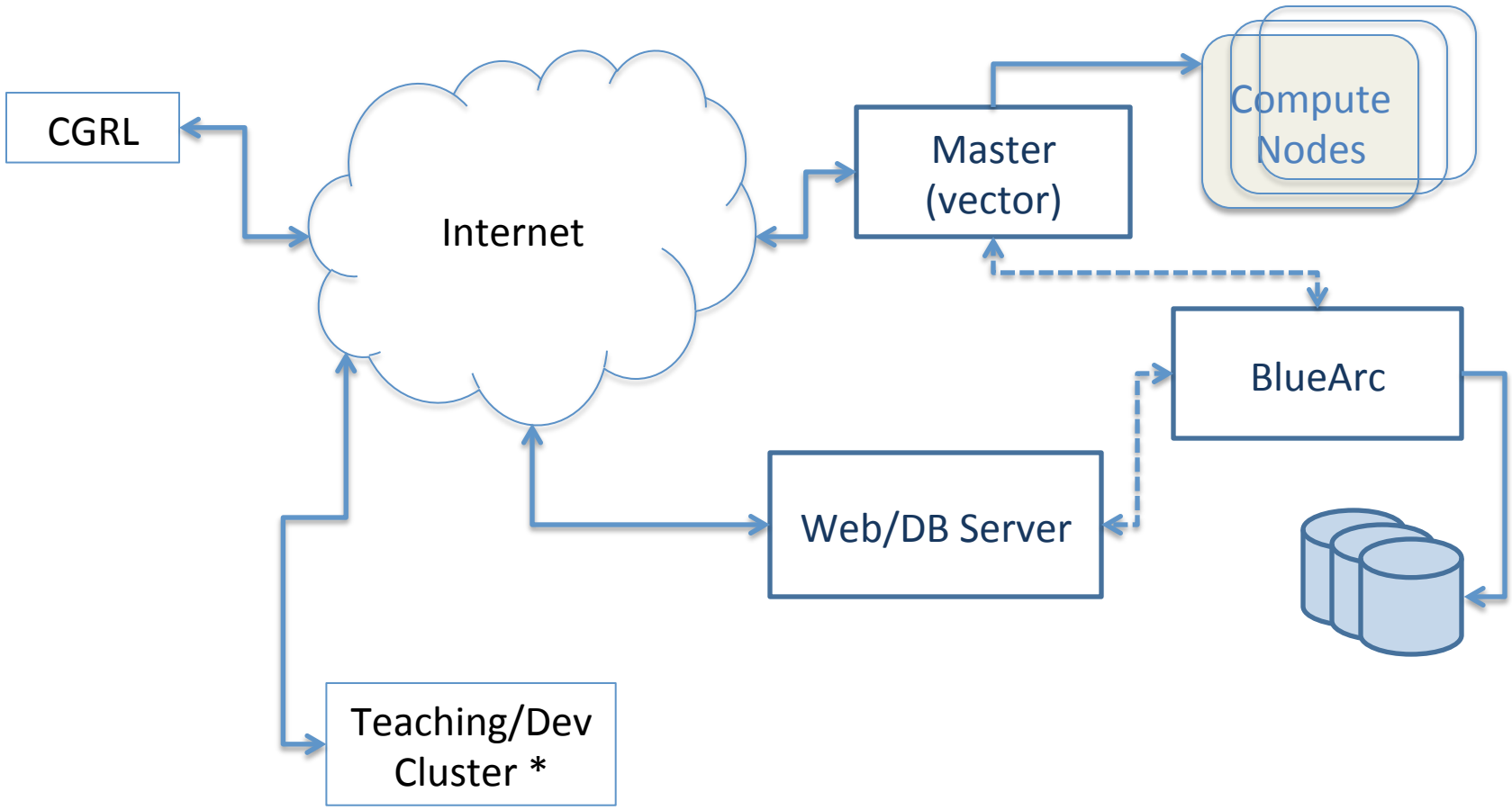
NGS Data Management

- Data Management
  - Sufficient storage during analysis and near-term storage
  - Plan for long-term storage and back-up
    - Offline hard drives – cheap option
    - Online raided NAS drives – better option
    - UC Storage and Back up (approx. $0.18/GB/mo.) – set up can be difficult, but no hardware to manage

Information Flow



CGRL Cluster Architecture



Computational Resources - Hardware

- Compute cluster running Linux (Centos 5.6)
  - Master node: Intel Xeon, 48GB memory (12 cores)
  - 4x Compute node: Intel Xeon
    - 96 GB memory, 12 core each (48 cores total)
  - 1x Compute node: AMD Opteron
    - 256 GB memory (48 cores)
  - Web server: Intel Xeon, 48GB memory (12 cores)
  - Bluearc storage server with >35TB usable storage
- PBS/Torque job queue management

Computational Resources - Software

- Analysis software for next-gen sequence
  - RNASeq, ChIPSeq, de novo assembly
- Reference data will be incorporated as needed
  - Genomes, other reference sequences
- Being implemented/evaluated
  - Browsing/visualization tools (Gbrowse, UCSC)
  - Galaxy workflow management tool

System Administration

- Colocated at IST data center (Warren Hall)
- Cluster managed by HPCS group from LBNL
- Crypto-card based access to cluster for better security (uses One-Time Passwords)

238 KH Facility

- Facility equipped for
  - Teaching, seminars, workshops
  - Interactive discussion area with large display
  - Workstations
    - Gigabit connection to campus network.
    - Self-managed back up of data
  - Monitors for data visualization
  - HD Projector

238 KH Facility - 2

- Workstations (2x)
  - 3.0 GHz, quad-core Intel processors
  - 12 GB, 1333 MHz memory
  - 30” Dell monitors
  - Windows 7 Professional, 64-bit
  - Linux Virtual Machine (Ubuntu 10.6, 64-bit)

238 KH Facility - 3

- Monitors (4x)
  - 30” Dell UltraSharp widescreen flat panel
  - 2560 x 1600 pixel resolution
  - DisplayPort, DVI-D, HDMI connections
    - Mini-DP converter available.

Funding for CGRL

- Initial infrastructure paid through grants from CACB, QB3, Bioscience Deans
- System administration & Maintenance
  - Recharge from CGRL users
- Upgrades and expansion as users increase
  - Compute nodes
  - Storage

Current CGRL Recharge Rates

- Annual cluster access fee per PI/group - $1500
  - Up to 3 users - login accounts, home directory (20 GB)
  - Large (few TBs) scratch/temp storage on cluster
  - Access to use CGRL facilities in 238 KH
  - Data storage for group on the cluster – 300 GB
- One time set-up cost for each user - $200
  - Crypto-card for compute cluster access
- Under consideration*
  - Hourly consultation rate (0.5hr blocks)
  - Usage-based cluster charge (beyond a threshold)

Fall 2012 Workshops Planned

- Genomic analysis at Unix command line
  - 10/10/2012 (registration online)
- DeNovo Assembly - Oct 22, 2012
- Genotyping & SNP calling – Oct 29, 2012
- RNA-Seq & ChIP-Seq - Pending

Other Meetings & Resources

- Repeating some workshops
  - Planning for Spring semester
- Making materials available online – CGRL Wiki
- Berkeley sequencing group meets here alternate Fridays at 10 am

Communication / Contact

- Websites URL
  - http://qb3.berkeley.edu/qb3/fgl
  - http://qb3.berkeley.edu/gsl
  - http://qb3.berkeley.edu/qb3/cgrl
- cgrl-announce @ lists.berkeley.edu
  - Anyone interested in related announcements
- CGRL Wiki: http://cgrlucb.wikispaces.com