Illumina Overview*

Adrian Reich

Brown University

August 2, 2010

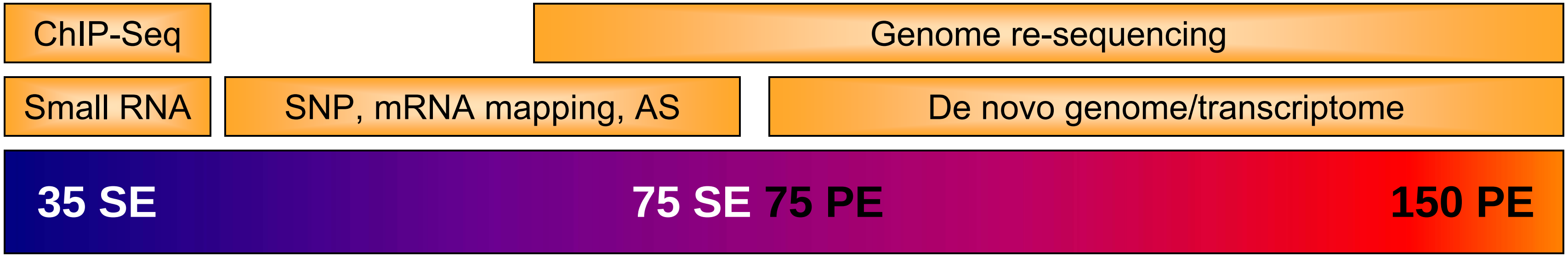
*Not definitive, but this should get you started…

Experimental Design on Illumina GAIIx

- Possible uses for the Genome Analyzer IIx (GAIIx)
  - Genomic DNA (promoters, ORF’s, non-coding RNA’s, etc.)
    - Re-sequencing, de novo assembly
    - Illumina genomic sequencing data sheet
    - De novo assembly technical note
  - mRNA-Seq (expressed transcripts)
    - SNP’s, alternative splicing (AS), mapping, de novo assembly
  - ChIP-Seq (identify sites of protein DNA interaction)
    - Illumina ChIP-Seq sequencing data sheet
  - Small RNA (identification of piRNA, miRNA, siRNA, etc.)
    - Illumina small RNA sequencing data sheet

More sequencing is not always better

- Different experiments require different amounts of sequencing effort
- To sequence one genome may require the same effort as two to twenty transcriptomes depending on genome size
  - Repository for all known genome sizes
- Each lane of a GAIIx will yield 30-40 million sequenced reads
- Each read is a user specified length of 35bp Single End (SE) up to 150bp Paired End (PE) which yields 1.25Gb to 10.8Gb per lane, respectively
  - Genomes – Reads should be sequenced longer, preferably PE
  - mRNA-Seq – Most variable in read length based on application, mapping needs short reads, de novo needs long PE reads
  - ChIP-Seq – 35bp is all that is needed for definitive mapping to genome for most experiments
  - Small RNA – By definition they are small and therefore do not require long reads



ChIP-Seq

Genome re-sequencing

Small RNA

SNP, mRNA mapping, AS

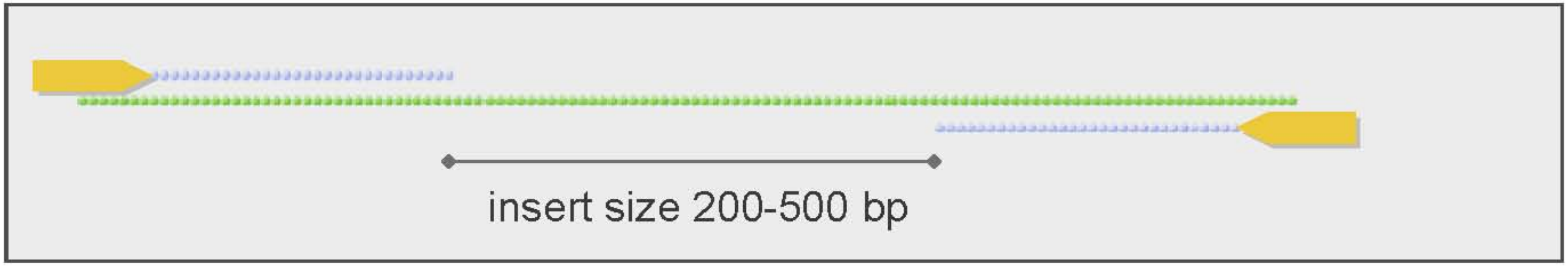
De novo genome/transcriptome

35 SE

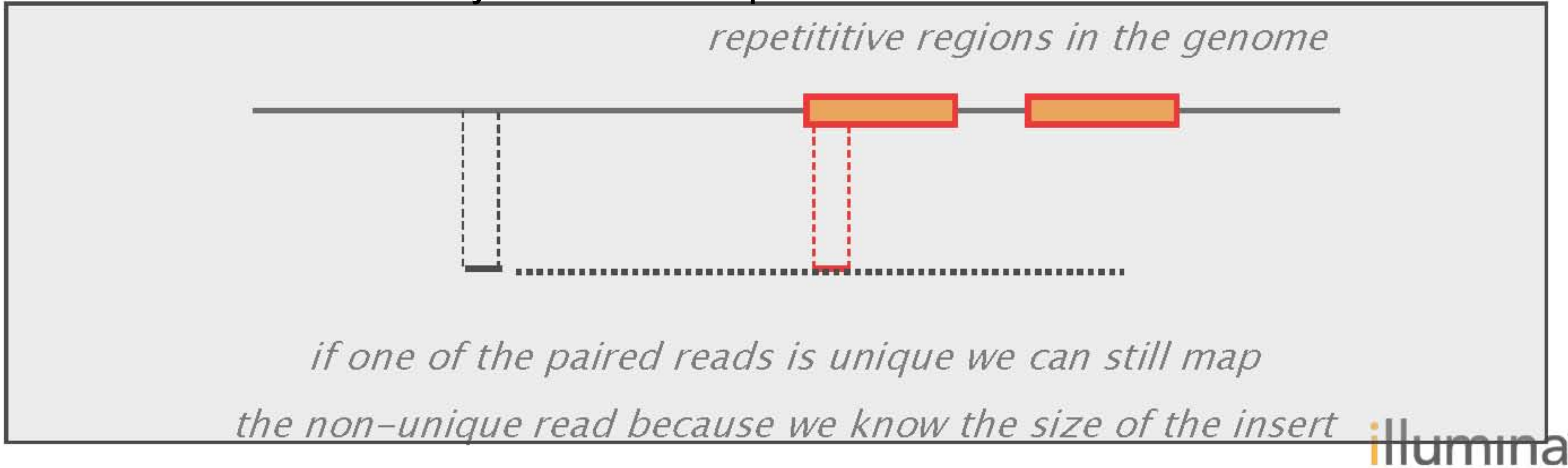
75 SE 75 PE

150 PE

Paired End – Sequencing Both Ends

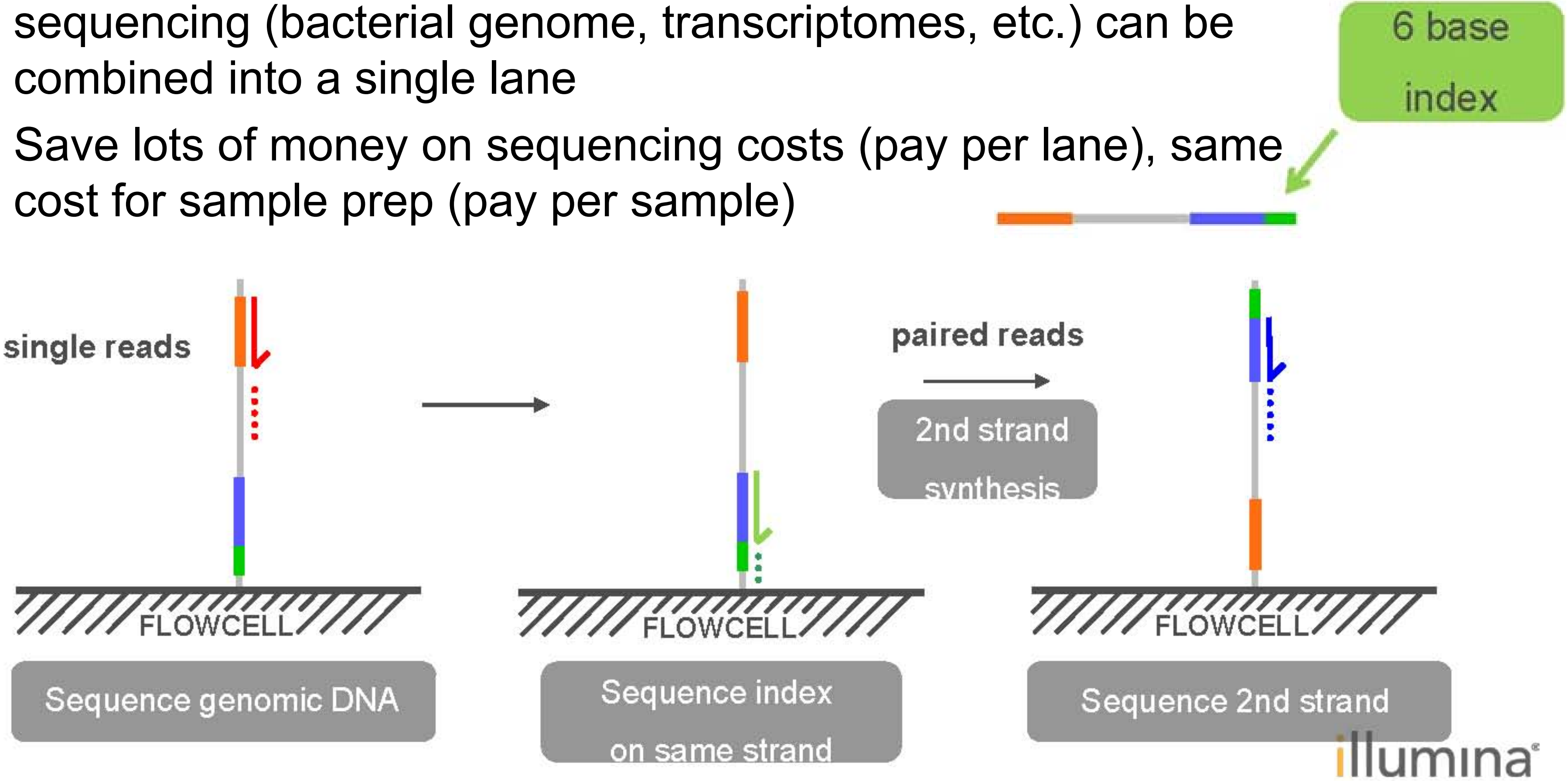


- Paired end sequencing at length X yields much more information than 2 SE reads at length X – Single end vs. paired end vs. mate pair
- Although the sequence is unknown between the 2 reads, the reads are known to be linked by a defined sequence


repetititive regions in the genome
if one of the paired reads is unique we can still map
the non–unique read because we know the size of the insert

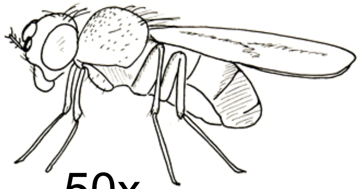
Multiplexing – Split Sequencing Effort

- Multiplex up to 12 samples per lane, each sample has a 6bp barcode – Illumina multiplexing datasheet
- Samples that are over sampled on one lane of GAIIx sequencing (bacterial genome, transcriptomes, etc.) can be combined into a single lane
- Save lots of money on sequencing costs (pay per lane), same cost for sample prep (pay per sample)

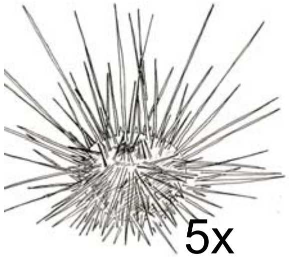


Planning the Sequencing Run

Read length	36 bp	76 bp	100 bp
Number of clusters	192M-240M (200-250K/tile) 300M maximum (~300K/tile)		
Gigabases Single-read / run time	5 – 9 Gb 2 days	12 – 15 Gb 3.5 days	16 – 24 Gb 5 days
Gigabases paired-end reads / run time	11 – 18 Gb 4 days	24 – 30 Gb 8.0 days	32 – 48 Gb 10 days
Avg. raw accuracy	99.25%	98.5%	98.0%

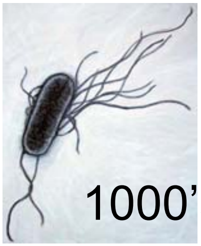


50x



5x

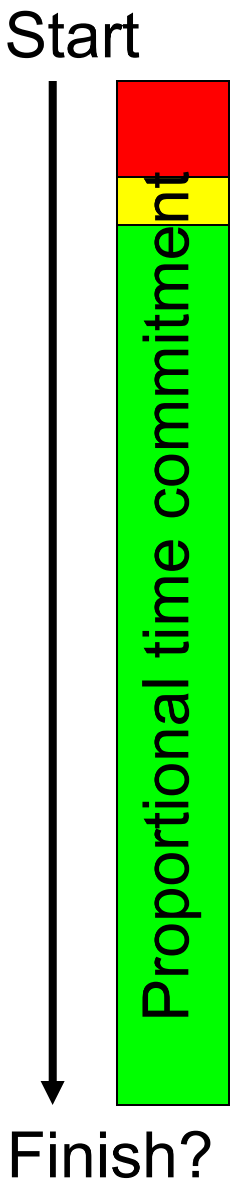
2x



1000’sx

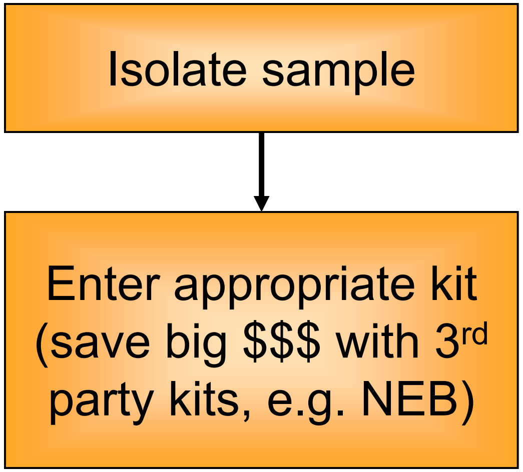
Fold coverage of genome per lane of GAIIx at 76bp PE

Three Steps of Illumina Sequencing



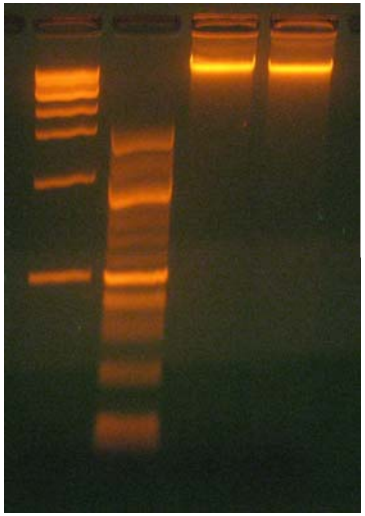
1. Sample Prep (2-21 days)
   - Library construction, use Illumina kits or 3rd party kits, (optimizations by Sanger Institute)
2. Cluster and Sequence (2-14 days)
   - Generate clusters and then sequence
3. Data Analysis (short to very long)
   - Illumina and/or 3rd party software depending on application
     - Actual computation time is very short
       - De novo genome assembly: hours-2 days
       - De novo transcriptome assembly: 2-3 days
       - ChIP-Seq mapping: hours-1 day
     - Figuring out which software is best for application and what to do with the data: weeks-months

Library Construction Workflow

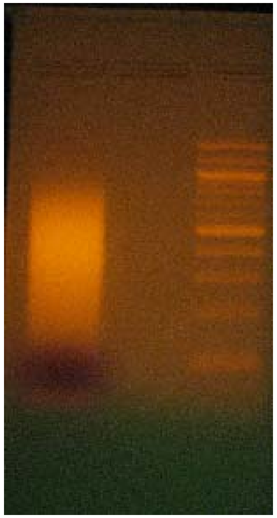

Isolate sample

Enter appropriate kit (save big $$$ with 3rd party kits, e.g. NEB)

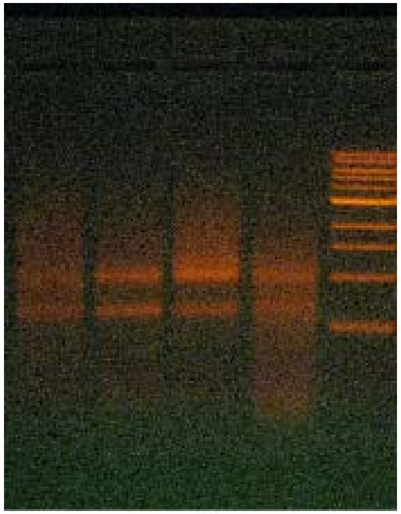
gDNA



ChIP

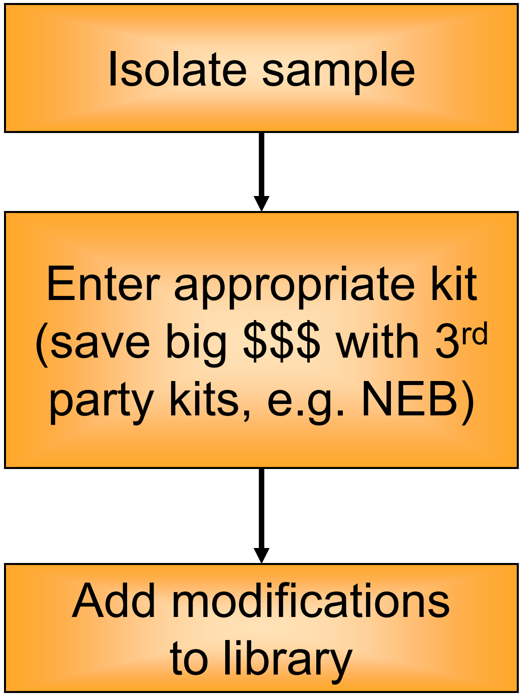


RNA



- Use high quality sample, garbage in = garbage out
- Start each kit with maximum recommended amount if possible
- Process sample (as per kit protocol) to generate dsDNA between 150bp and 500bp

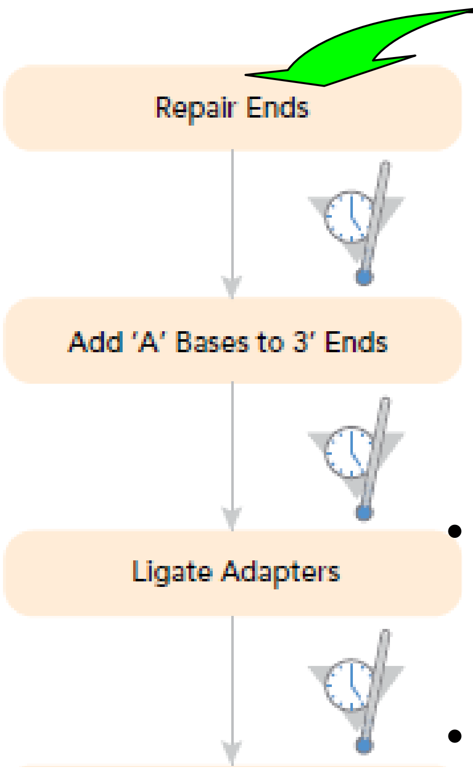
Library Construction Workflow

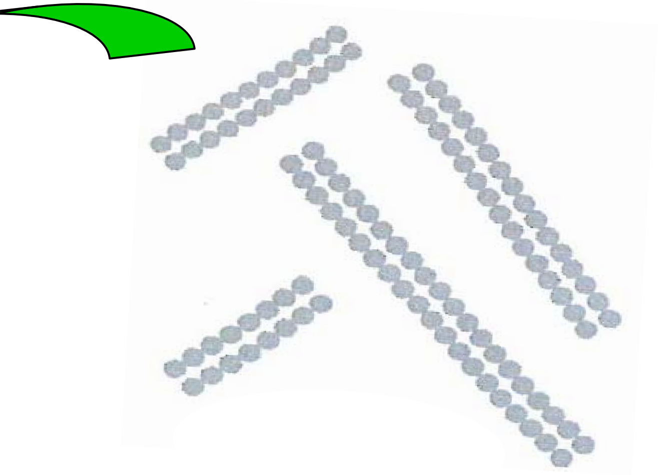


Isolate sample

Enter appropriate kit (save big $$$ with 3rd party kits, e.g. NEB)

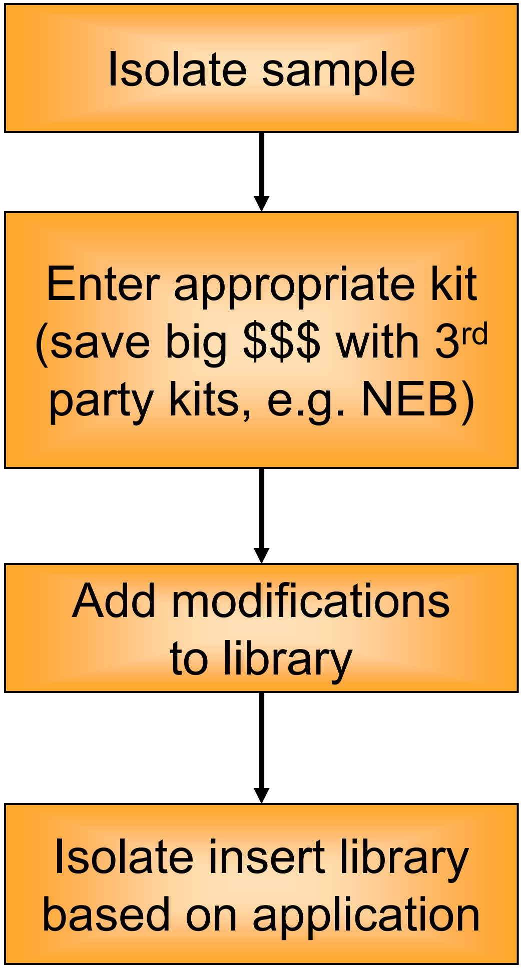
Add modifications to library





- All kits generate dsDNA then undergo the same reactions
- dsDNA is biochemically modified so each fragment has identical ends
- Adaptors have numerous proprietary modifications, some are known and can be purchased from outside vendors (IDT)

Library Construction Workflow

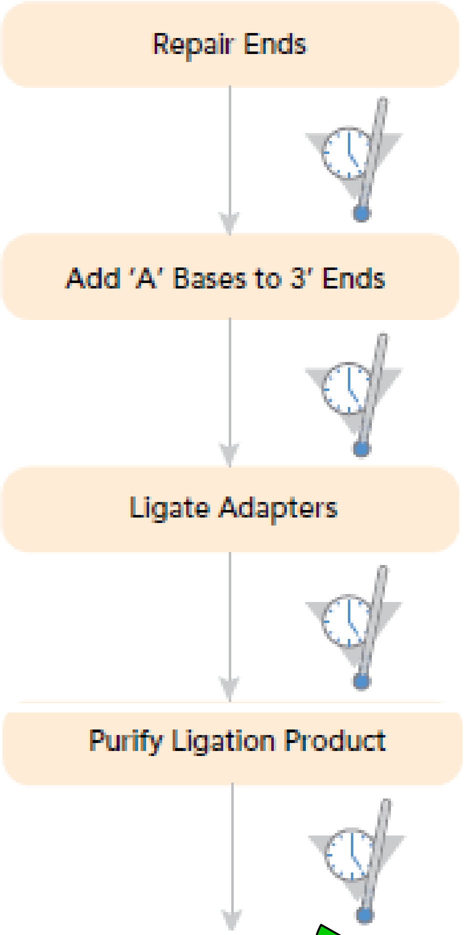


Isolate sample

Enter appropriate kit (save big $$$ with 3rd party kits, e.g. NEB)

Add modifications to library

Isolate insert library based on application



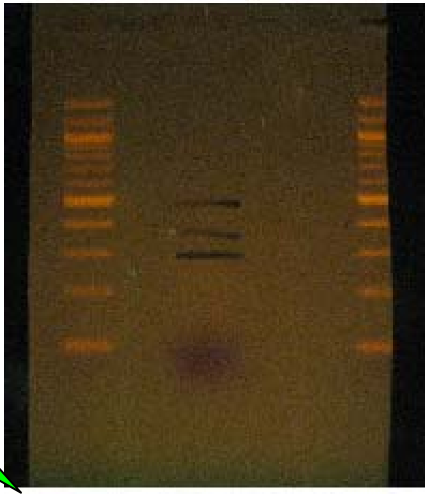
Repair Ends

Add 'A' Bases to 3' Ends

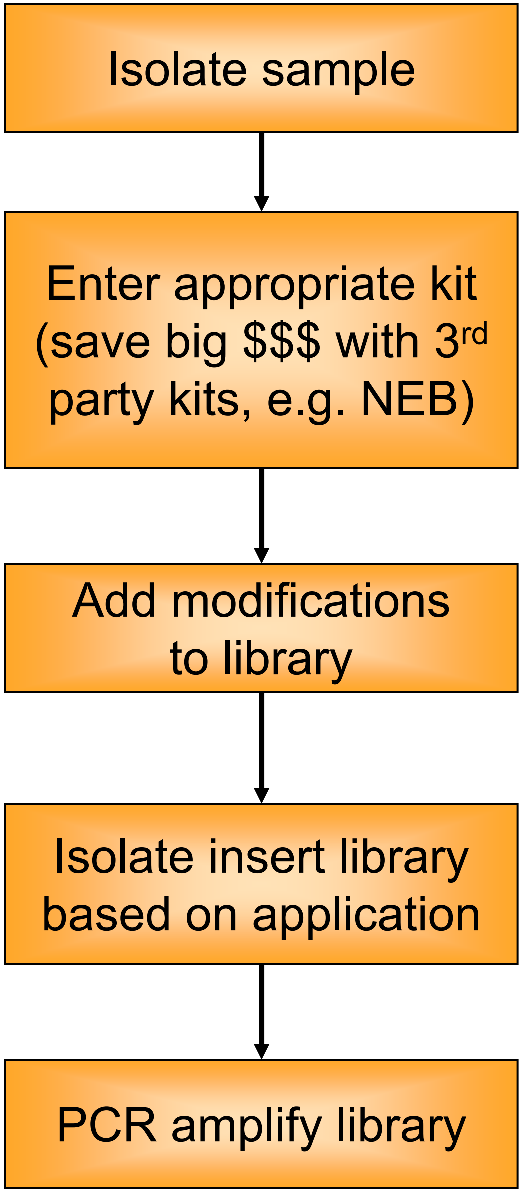
Ligate Adapters

Purify Ligation Product

- Different applications require different libraries
  - De novo sequencing needs very tight bands, ChIP-Seq needs a smear
- Multiple libraries (with different insert sizes) can be prepared from the same sample (here 3)
- Some sample preps require the cutting of invisible bands



Library Construction Workflow



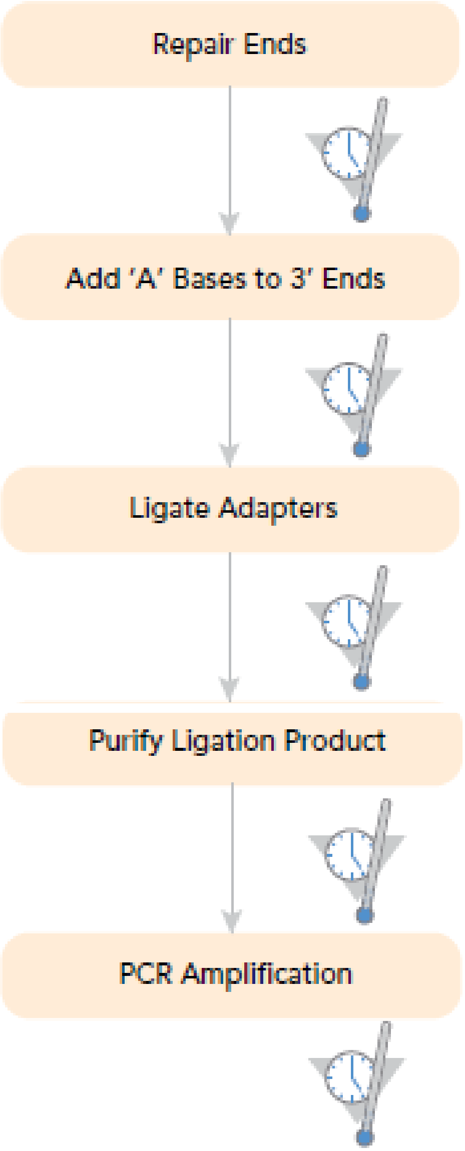
Isolate sample

Enter appropriate kit (save big $$$ with 3rd party kits, e.g. NEB)

Add modifications to library

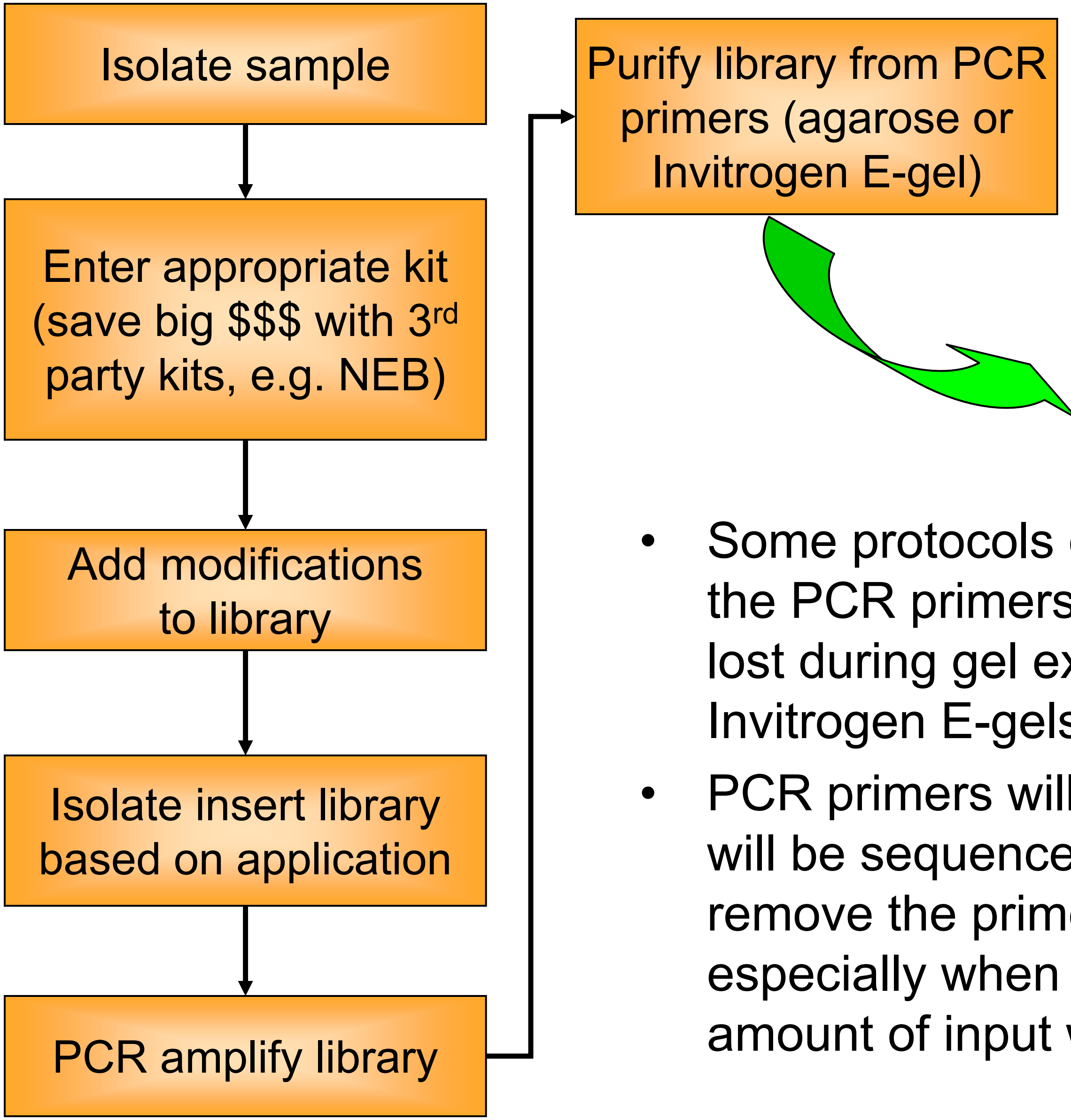
Isolate insert library based on application

PCR amplify library



- PCR amplification enriches for specific library sizes
- The PCR step will add additional sequence to the library (described here)
- The primers also have some proprietary modifications, but can also be purchased from outside vendors (IDT)

Library Construction Workflow



Isolate sample

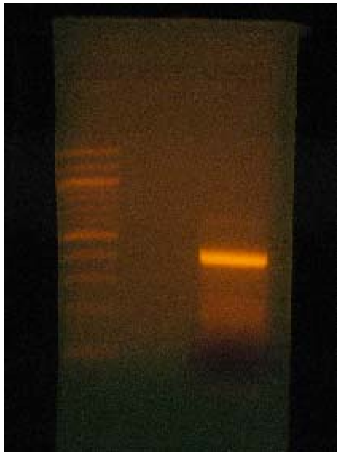
Enter appropriate kit (save big $$$ with 3rd party kits, e.g. NEB)

Add modifications to library

Isolate insert library based on application

PCR amplify library

Purify library from PCR primers (agarose or Invitrogen E-gel)



- Some protocols do not include the removal of the PCR primers because too much sample is lost during gel extraction: in these cases use Invitrogen E-gels, the total yield is much higher
- PCR primers will cluster on the cBOT and they will be sequenced, it is very important to remove the primers for library quantification, especially when less then the maximum amount of input was used

Library Construction Workflow

Isolate sample

Enter appropriate kit (save big $$$ with 3rd party kits, e.g. NEB)

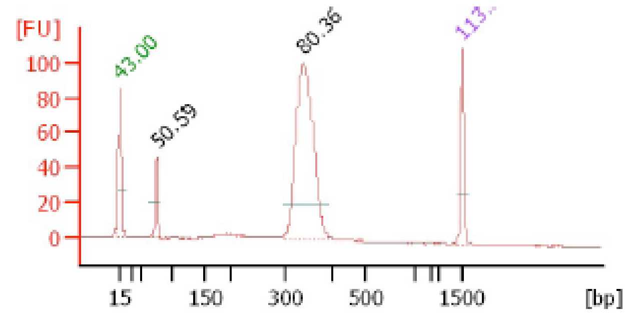
Add modifications to library

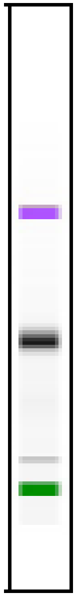
Isolate insert library based on application

PCR amplify library

Purify library from PCR primers (agarose or Invitrogen E-gel)

Run Bioanalyzer



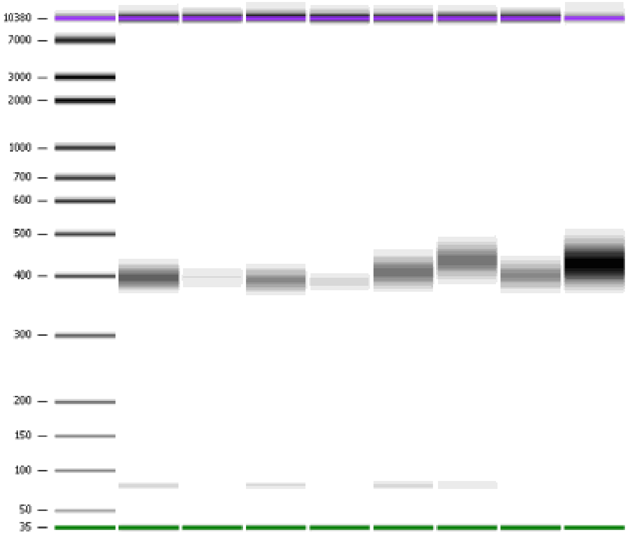


Overall Results for sample 4 :

Number of peaks found: 2

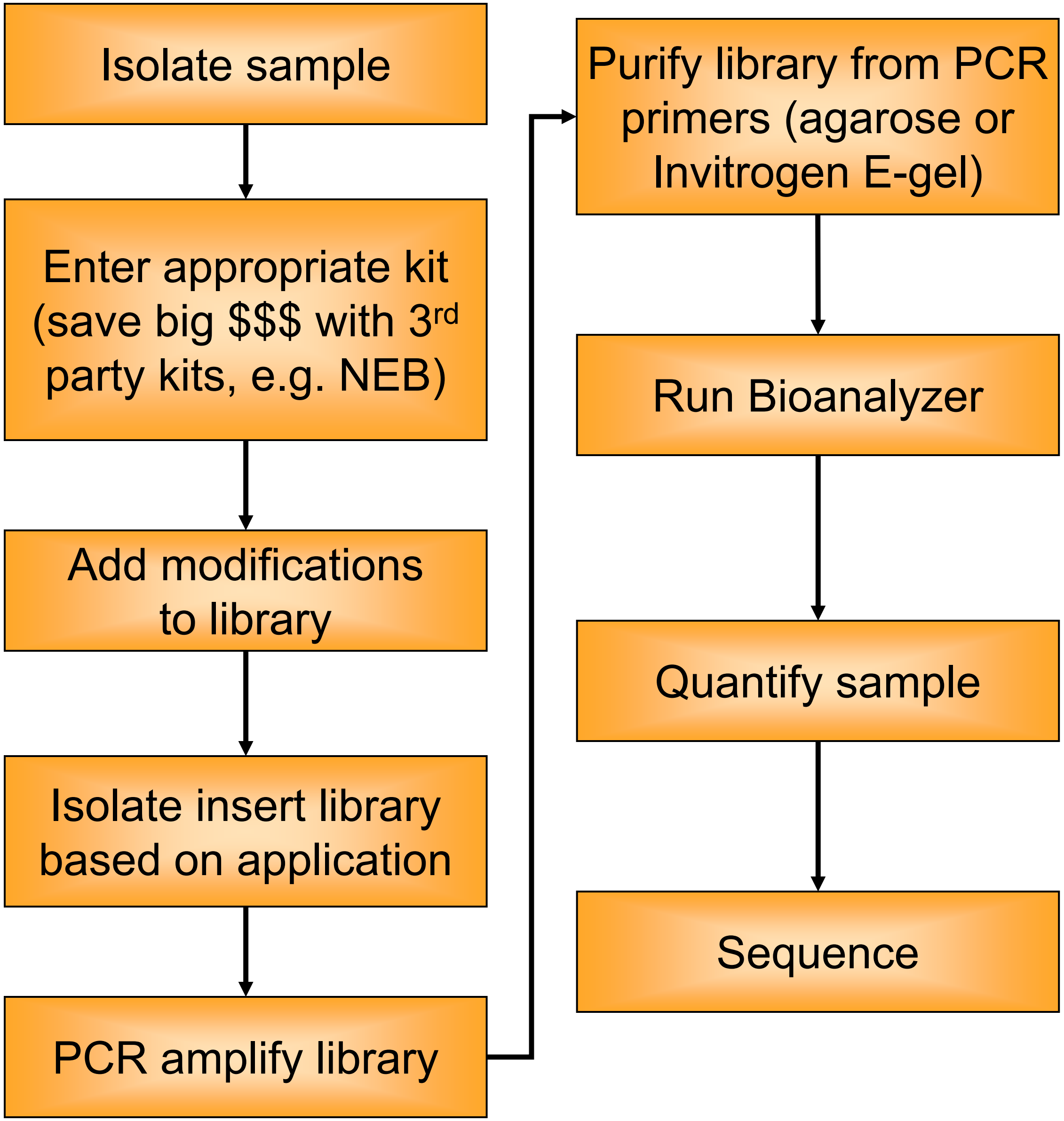
Peak table for sample 4 :

Peak		Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	◀	15	4.20	424.2	Lower Marker
2		74	1.75	35.8	
3		338	14.91	66.9	
4	▶	1,500	2.10	2.1	Upper Marker



- After library prep is complete, 1µL of the prep will yield a measurement of all the bands in the sample and the specific concentration for each

Library Construction Workflow



Isolate sample

Enter appropriate kit (save big $$$ with 3rd party kits, e.g. NEB)

Add modifications to library

Isolate insert library based on application

PCR amplify library

Purify library from PCR primers (agarose or Invitrogen E-gel)

Run Bioanalyzer

Quantify sample

Sequence

- In order to know how much library to load on the sequencing machine an accurate quantification of the library is necessary
- Four methods are available (all four are found in the core facility), qPCR is best but if this method is used, the PCR primers from library amplification must be removed

• Nanodrop	–
• Qubit	+
• Bioanalyzer	+
• qPCR	+++

Illumina Library Quantification by qPCR

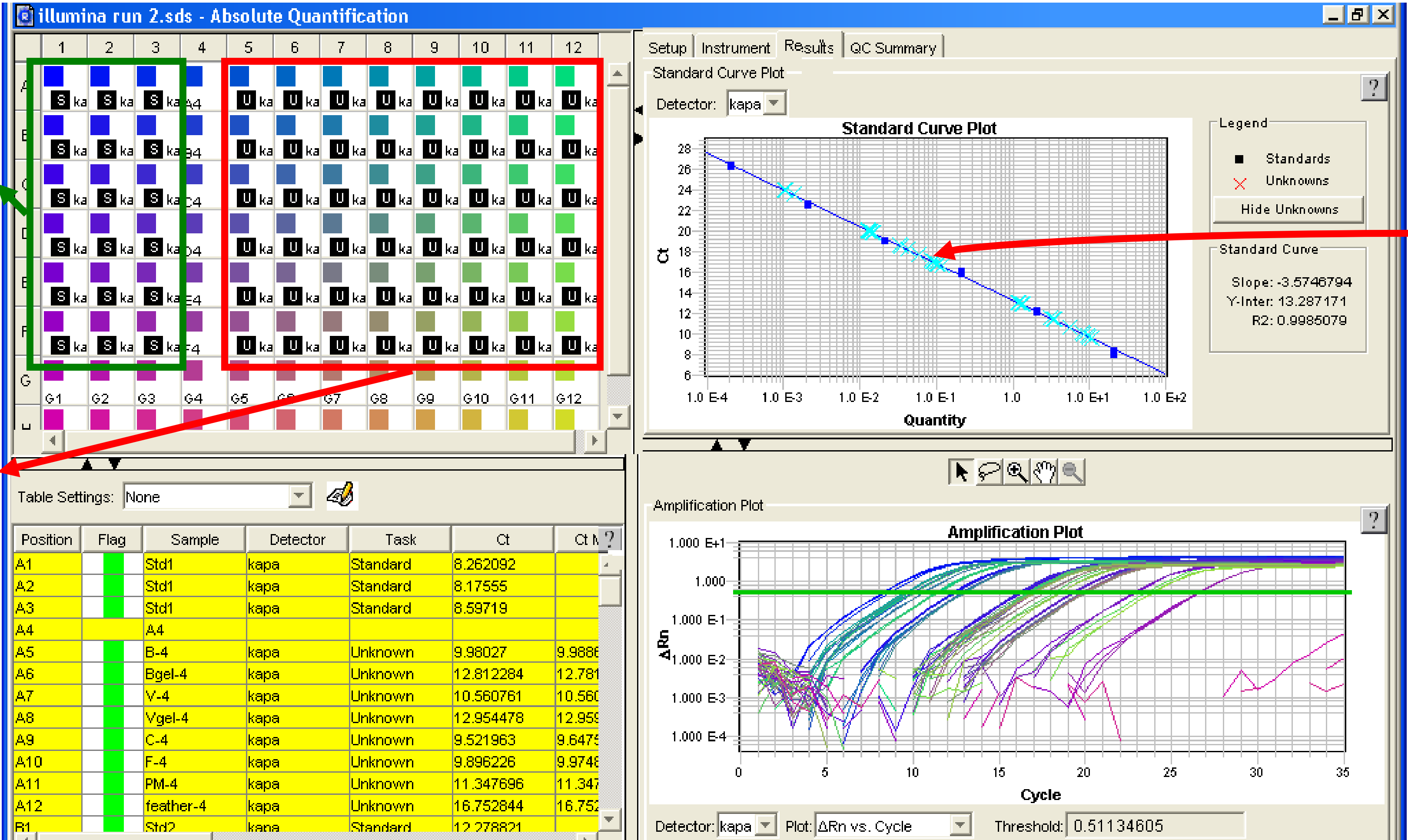
KAPABIOSYSTEMS – Evolved enzymes for 500bp qPCR

- Unknown libraries are compared to standard curve drawn from known DNA standards included in kit
- Samples are run in triplicate resulting in high accuracy (Kapa datasheet)

Known standards

Unknown libraries

Unknown libraries plotted on standard curve



Slope: -3.5746794
Y-Inter: 13.287171
R2: 0.9985079

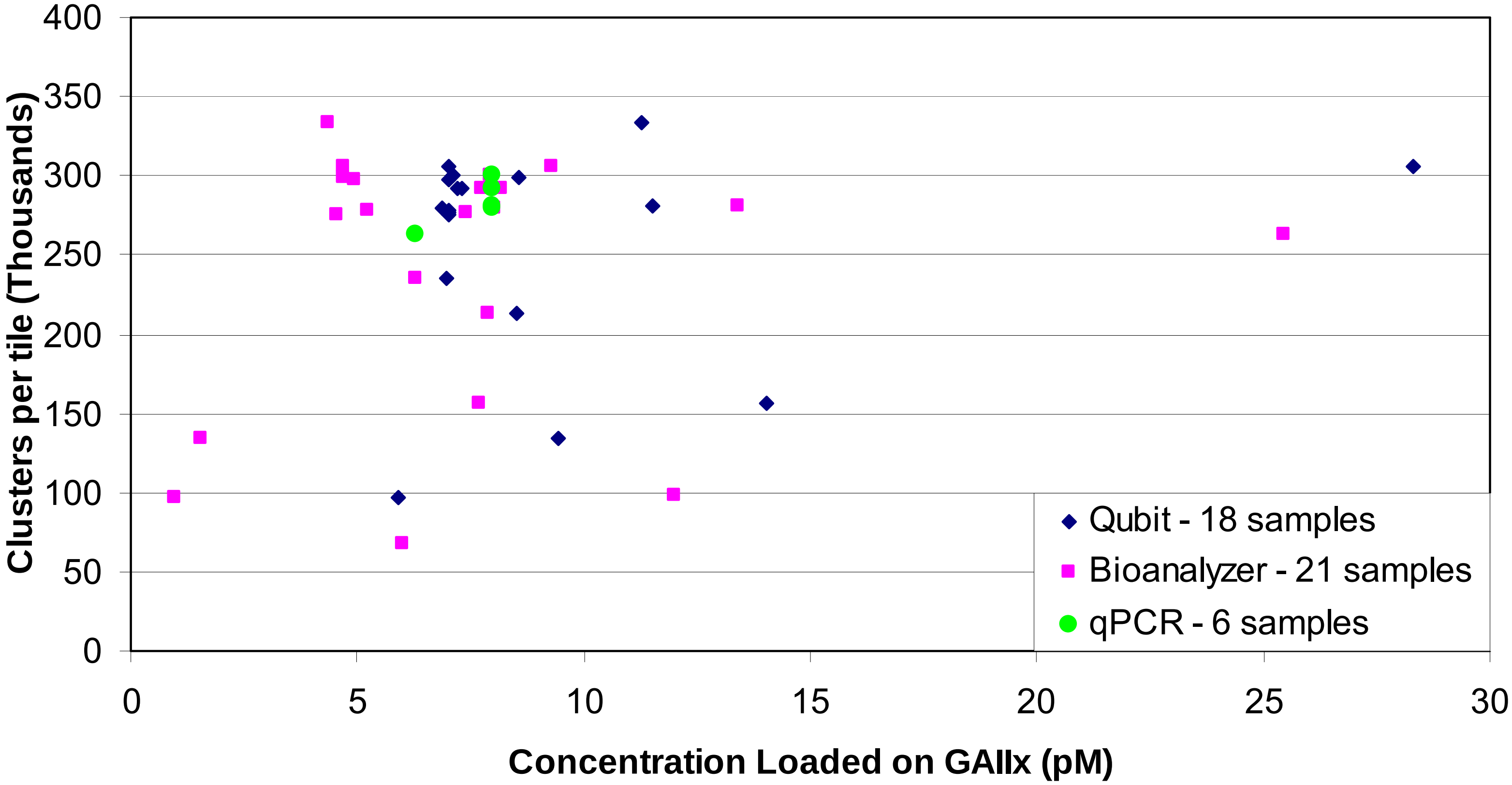
Threshold: 0.51134605

Table Settings: None

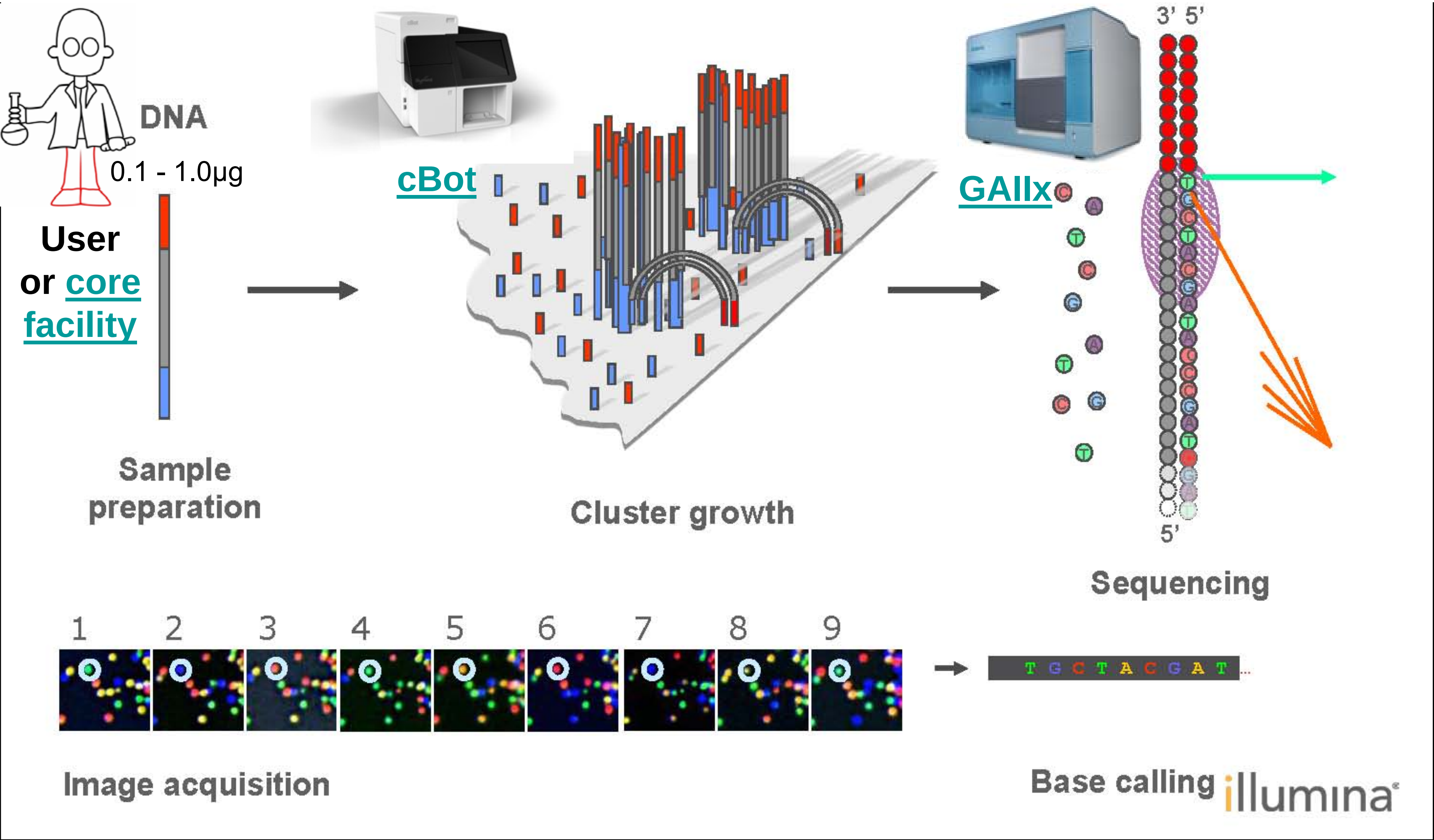
Position	Flag	Sample	Detector	Task	Ct	Ct M
A1	🟩	Std1	kapa	Standard	8.262092	
A2	🟩	Std1	kapa	Standard	8.17555	
A3	🟩	Std1	kapa	Standard	8.59719	
A4		A4				
A5	🟩	B-4	kapa	Unknown	9.98027	9.9886
A6	🟩	Bgel-4	kapa	Unknown	12.812284	12.781
A7	🟩	V-4	kapa	Unknown	10.560761	10.560
A8	🟩	Vgel-4	kapa	Unknown	12.954478	12.959
A9	🟩	C-4	kapa	Unknown	9.521963	9.6475
A10	🟩	F-4	kapa	Unknown	9.896226	9.9748
A11	🟩	PM-4	kapa	Unknown	11.347696	11.347
A12	🟩	feather-4	kapa	Unknown	16.752844	16.752
B1	🟩	Std2	kapa	Standard	12.278821	

Sample Prep

- Comparing the three methods of library quantification over 21 samples run at Brown, qPCR is clearly the best predictor of cluster density
- The qPCR samples had primers removed after the PCR amplification step
- 300K clusters per tile (120 tiles/lane = 36 million reads) is an optimally loaded lane, lower density results in less data, higher density can result in zero data, hence an accurate library quantification method is important



Sequencing on the GAIIx



User or core facility

0.1 - 1.0µg

cBot

GAIIx

Post sequencing Data Analysis

- Run data is posted to the Brown University Illumina sequencing webpage
- The raw data is transferred to the Oscar computing cluster, each lane will generate raw sequence files ranging in size from 1GB to 20GB
- To manipulate the files on Oscar you must get an account, large users are encouraged to get Priority or Condominium accounts
- Illumina supports some types of sequencing applications with analysis software, de novo assembly of genomes or transcriptomes are not supported (though mapping is) and so 3rd party software must be used
- Contacts:
  - Brown Core Facility
    - Christoph Schorl
    - Hilary Hartlaub
    - Marissa Kielbasinski
  - Bioinformatics Analyst
    - Lingsheng Dong
  - CCV
  - Illumina Technical Support
    - (800) 809-4566

Happy Sequencing!