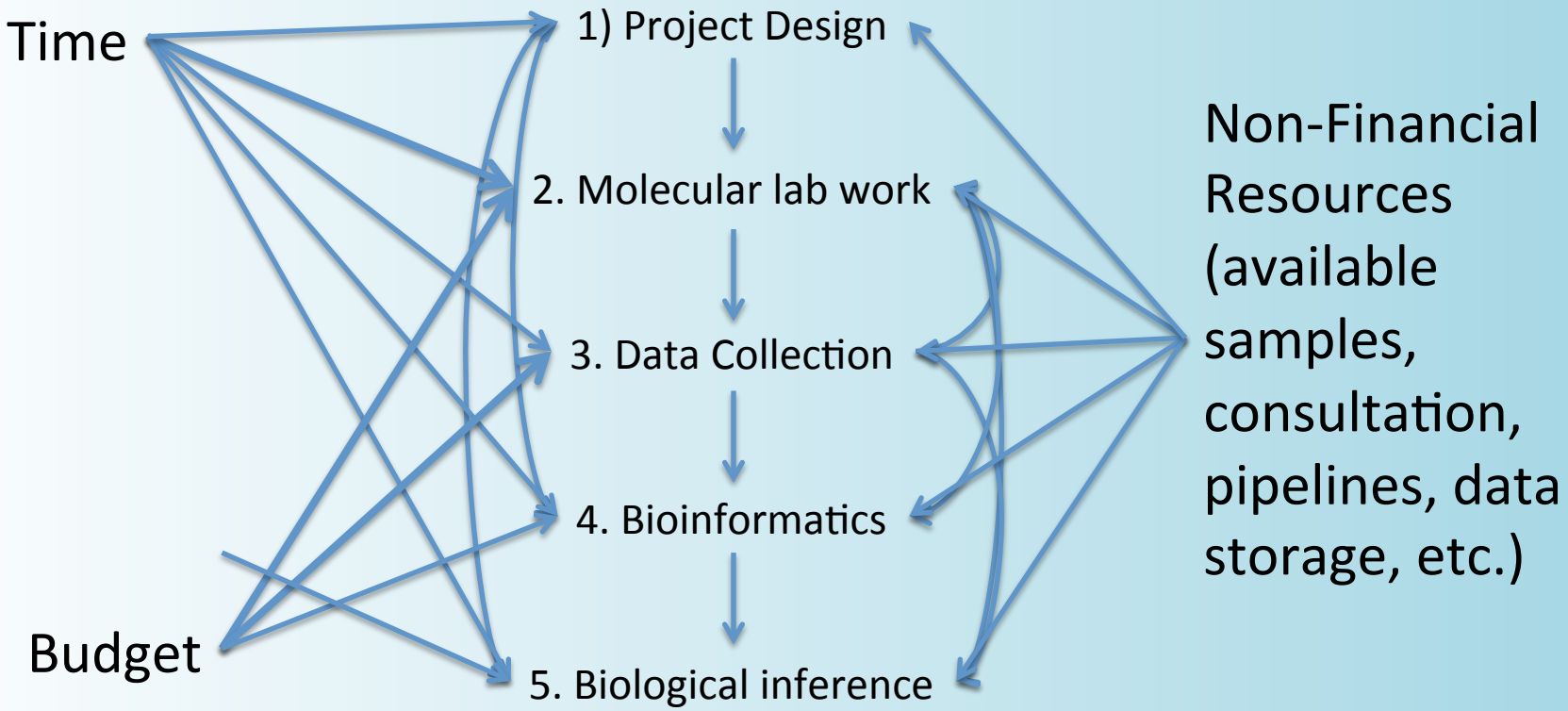
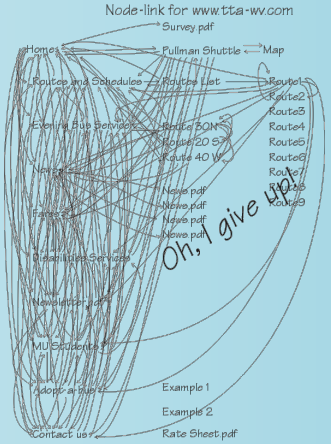
Getting Started in the Wet Lab with High-Throughput Sequencing Projects: best practices & planning ahead

Lydia Smith
Manager, Evolutionary Genetics Laboratory,
Museum of Vertebrate Zoology
lydsmith@berkeley.edu
11 February 2016

Easy steps to HTS project success



Confusogram



- But don’t despair!
- Once you begin thinking about what you really need from your data and what your limitations are in terms of time, budget, and resources, options will disappear and decisions will become easier.
- **The two uppermost parts of the process (Project Design & Molecular Work) determine everything that follows.** Samples can be easily resequenced (although for a high cost) and reanalyzed (although at a cost of time), but if libraries were not made with care, made in the right format, and made to target the most useful type of data, you may not be able to get the data you need from the project without starting over.

Problems to consider

- How much data is required to answer biologically study relevant questions?
- What type of data is most informative?
- How many samples are needed in the study?
- Project budget
- Number of individuals to sequence
- Number of markers to target
- Availability of reference genome
- Coverage needed for downstream analysis

Importance of Careful Study Design

It is tempting to dive in and just collect a lot of data, but if these questions aren't thoughtfully considered before starting, the desired bioinformatic approaches and biological inferences may not be possible and large amounts of money can be wasted.

Consult with someone informed as early as possible before you begin lab work: PI, labmate, collaborators, CGRL

http://qb3.berkeley.edu/qb3/cgrl/contact.cfm

Campus Resources for high-throughput sequencing: QB3 Labs

- Computational Genomics Resource Laboratory (CGRL):
  - project design & preliminary bioinformatics discussions prior to start of a project
  - computer cluster and pipelines designed for large-scale genomic data

- Functional Genomics Laboratory (FGL)
  - will perform library preparations. This is expensive per prep (> $100 genomic DNA, > $300 RNA) but a great option for projects with very large budgets or only a few libraries to make.
  - has specialized pay-per-use equipment for researchers making their own libraries: Covaris, Bioanalyzer, Pippin Prep

- Vincent Coates Genomics Sequencing Laboratory (GSL)
  - supports Illumina sequencing technology on campus
  - PacBio library preparations (instrument at UC Davis)

About the Evolutionary Genetics Laboratory

- We are not a fee-for-service lab like the ones previously mentioned. Instead we offer training and trouble-shooting support, access to specialized equipment (bioruptor sonicator, plate magnets, hybridization equipment), bench space, and a storeroom with essential reagents and adapters in stock in order for researchers to perform their own wet-lab work.
- Molecular biology facility for the Museum of Vertebrate Zoology, but we also allow lab membership for anyone in the wider Berkeley community so long as they are:
  - working on project types that we support (especially, but not limited to, genomic DNA capture by hybridization, transcriptome/RNA-Seq, RADseq);
  - willing to put up with some of the challenges of a shared lab environment as well as reaping the benefits (**patience** is key) and to act as a member of a lab community;
  - able to complete the full orientation and safety training for our facility;
  - in agreement to purchase the bulk of their project supplies through our storeroom (the overhead percentage funds equipment, and salaries; for common items, our discount for bulk purchases is greater than the overhead.)

About the Evolutionary Genetics Laboratory: Supported Techniques

Fully supported (all supplies in stock except custom probes):

- genomic DNA library preparation with PCR (reagents: $10-15/prep)
- RNA library (poly-A selection) (reagents ~$50/prep)
- Hybridization captures (Nimblegen, MyBaits) (reagents beyond library preps & probes ~$100/capture)

Partially supported (can work with you to get additional items as needed)

- genomic DNA library, PCR-free
- RNA library (rRNA depletion)
- RAD-Seq library
- Amplicon sequencing

About the Evolutionary Genetics Laboratory: Specialized Equipment

- Diagenode Bioruptor (sonicator)*
- Qubit (fluorometer)**
- Bioanalyzer**
- Magnetic plates
- MoBio PowerLyzer (bead beater)

*Covaris sonicator available in FGL

**Qubit & Bioanalyzer available in FGL & GSL

About the Evolutionary Genetics Laboratory: How to get involved:

- E-mail me with a brief description of your project, your past experience, and your time-table for completing the work: lydsmith@berkeley.edu
- We will discuss whether your project is a good fit for our resources and expertise
- We can schedule orientation and training sessions when time allows
- Payment for storeroom purchases: UC Berkeley chartstring or invoice information if funding is off-campus

Sequencing Machines

- What machines are available? (on-campus, elsewhere)
- What are your colleagues using?
- Do you need long reads?
- Do you need paired-end reads?
- Do you need high coverage?
- Do you need multiplexing?

Sequencing Machines 2016
UC Berkeley-supported technologies

Machine	Best Cost ($/MB)	Minimum run buy-in	Reads per run (millions)	Read Length	Paired-end?	Multiplexing?	Final Error Rate
ABI 3730 Sanger sequencing	$1500	$4	0.000001	Up to 1000 bp	yes	no	0.1-1%
Illumina HiSeq 4000	$0.03	$1400-$2300	300 (SR)	Up to 150 bp	yes	yes	0.1%
Illumina HiSeq 2500	$0.05	$950-$3730	120 (SR)	Up to 250 bp	yes	yes	0.1%
Illumina MiSeq	$0.11	$1100-$1750	25 (SR)	Up to 300 bp	yes	yes	0.1%
PacBio RS II (at UC Davis)	$0.50	$900	0.1	~10K*	no	yes (in 2015)	1-13%*

Updated from: http://www.molecularecologist.com/next-gen-fieldguide-2014/

Types of High-Throughput Sequencing Instruments

Short-read:

Illumina (huge market share, best cost per Mb, short reads only but growing—up to 300bp)

Life Technologies PGM/ProtonTorrent

Long Read/Single Molecule:

PacBio (over 10K): whole genome sequencing, full-length cDNA

Oxford Nanopore (still in beta): doi:10.1016/j.bdq.2015.02.001

"Synthetic" Long Reads using short-read sequencers:

Moleculo: Illumina library method of simulating long reads (10kb) with tagged Illumina data. User-purchased kit (very complicated)

10X Genomics: long DNA molecules (50kb) partitions and prepares sequencing libraries in parallel such that all fragments produced within a partition share a common barcode. Requires special machinery (genome core)

For more options and details, see: CGRL Genome Assembly Workshop Feb 22 & 23:
https://cgrlucb.wikispaces.com/file/view/genomics_workshop_UCB_2015.pdf/564276201/genomics_workshop_UCB_2015.pdf

Local UC Sequencing facilities

QB3 Vincent Coates Genome Sequencing Laboratory:
http://qb3.berkeley.edu/qb3/gsl/docs/GSLRates2016.pdf
Director: shana.mcdevitt@berkeley.edu
Illumina HiSeq4000, HiSeq2500, MiSeq

UC Davis: http://dnatech.genomecenter.ucdavis.edu/prices/
Lutz Froenicke: lfroenicke@ucdavis.edu
Illumina (similar options to Berkeley), PacBio

(All UC customers receive internal rates at both facilities)

Whole Genome Sequencing

A single genome or transcriptome sequence is rarely the end goal of a scientific project.

Most projects require comparisons between multiple samples to answer the biological questions of interest.

This can be accomplished with WGS if

1) the genome size is small, and/or
2) there is a well-annotated reference genome, and/or
3) only low levels of coverage are required

Sequencing whole genomes has some drawbacks (cost, data storage, complexities of analysis), but it allows researchers to look comprehensively at the entire genome (well, duh), including promotors, regulatory elements, and introns, not just the small slice of the genome in the other methods we will discuss.

When looking for signatures of selection and patterns of molecular evolution, a WGS will give the study a much greater chance of success than exome sequencing alone.

Even when genomic resources exist and coverage can be low, only a few large genomes can be run on a single lane so sequencing costs can be expensive.

Whole Genome Sequencing: what if we don't have a reference genome

De novo sequencing of full genomes (without pre-existing reference sequences) is becoming a real possibility for (relatively) moderate amounts of money.

But bioinformatics challenges remain for de novo assembly especially on large, highly repetitive, and low complexity genomes.

WGS not an appropriate strategy for comparative studies of large genomes unless there is already a well-annotated high-quality reference genome

However, sometimes a genome must first be sequenced in order to obtain the information needed for comparative methods.

Comparative Genomics: genome partitioning methods

If the size of the genome is small, whole genome sequencing of many samples is a real possibility to compare.

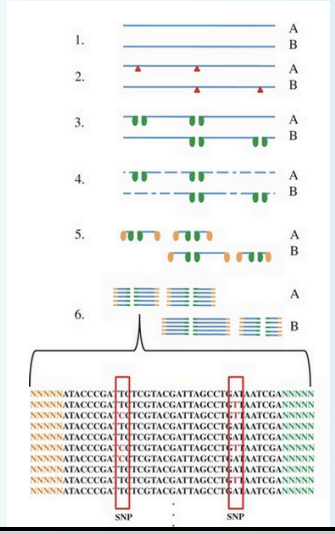
Otherwise, we must find ways to 1) reduce the amount of sequence collected while 2) ensuring that we obtain orthologous parts of the genome from all the samples and 3) ensuring that we collect the type of genomic data to allow us to address biologically meaningful questions

Common methods involve different types of library preparations:

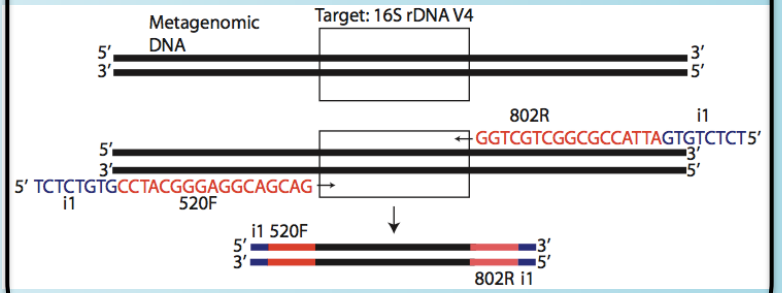
- Targeted capture of genomic DNA (often exons, but could be any desired region of the genome)
- Transcriptome sequencing (sequences of exons, differential gene expression)
- RAD-Seq (sequences of SNPs near restriction enzyme cut sites)
- ChIP-Seq (sequences of DNA at and near protein binding sites)
- Amplicon sequencing (sequences of multiple organisms in one or more PCR amplicons: metagenetics)

Comparative Genomics Tools: strategy depends on the research question

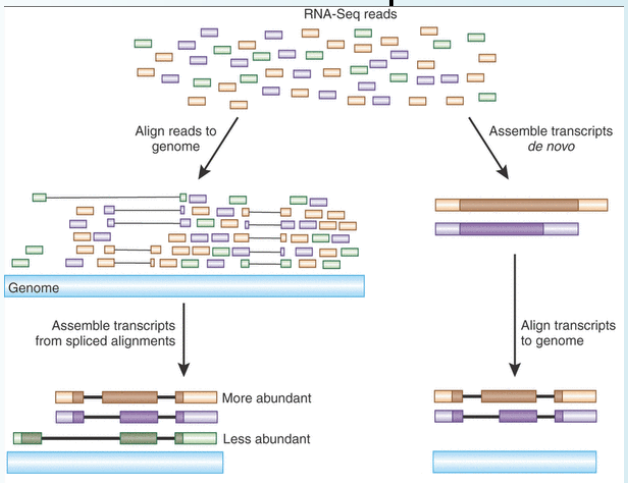
RAD-tag Sequencing



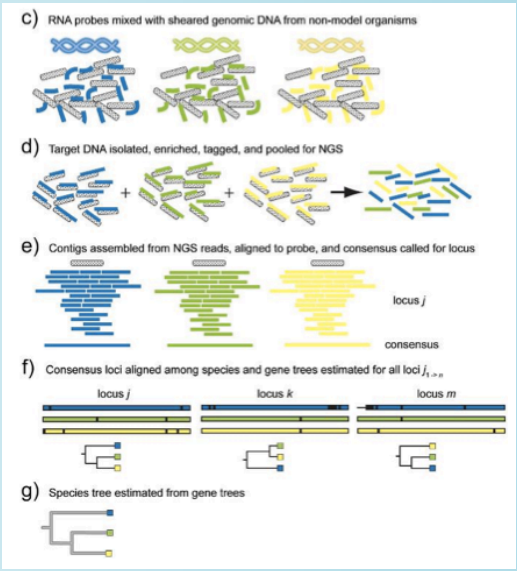
Amplicon Sequencing



RNA Seq



Targeted Enrichment



Targeted capture by hybridization

- Great approach for getting consistent sequencing results from the same genomic partition from multiple samples: population genetic & phylogenetic studies
- Can multiplex dozens of samples per lane, depending on target size
- Can be used with degraded samples
- Can target “high-value genomic regions”

- Usually requires a reference genome or transcriptome: *a priori* knowledge of target sequences in order to design custom probes (some predesigned kits are available). If unavailable, preliminary sequencing costs of transcriptome(s) or low coverage WGS will be additional cost/time
- Potentially expensive projects: library preparation costs are more than with RAD-Seq; additional probe synthesis and capture costs
- Challenging for organisms with large genome sizes and/or highly repetitive genomes

DOI: 10.1111/mec.13304

RNA-Seq

- Sequences the expressed transcripts in a tissue
- Data has more possible uses: can be used to get exon sequence information, transcript counts for differential gene expression studies, and/or alternative splicing information
- No previous genome or transcriptome information required

- Requires well-preserved tissue to extract high-quality RNA
- More expensive library preparation than DNA methods
- Highly expressed genes dominate the data; deep coverage required to find rare transcripts

http://rnaseq.agbioinfo.utk.edu/images/2/20/Transcriptome_Project_Design.pdf

Restriction-site Associated DNA markers: RAD-Seq

- Inexpensive way of identifying 100s or 1000s of SNPs in many closely-related organisms. Can gather large number of anonymous markers from many samples over a short period of time
- Widely used for inferring population structures, phylogeography, trait mapping, genetic maps, and association – plenty of case studies in the literature
- No pre-existing reference sequence needed

- Requires high quality DNA (although some hybrid approaches like HyRAD (http://dx.doi.org/10.1101/025551) and Rapture (DOI: 10.1534/genetics.115.183665) may allow integration of degraded samples)
- Collecting homologous markers from all samples is a struggle no matter how careful the planning and the lab work: struggle with locus drop-outs and high variance of depth across loci and individuals
- Poor choice for phylogenetics due to potential of mutations at RE cut-sites
- Challenging for organisms with large genome sizes and/or highly repetitive genomes

doi:10.1038/nrg.2015.28

Amplicon Sequencing

- Useful for metagenomics applications where there are multiple taxa present in each sample (e.g. 16S for investigating diversity and structure of complex microbial communities & populations)
- Cost-effective if # of loci is low and # of samples is high
- No reference genome or transcriptome information needed; just enough sequence to design primers from

- Not cost or time-effective for many loci unless emulsion PCR is used
- Requires a successful amplification for each sample
- Amplicon length is limited by sequencing read lengths; since the libraries are not sheared, information more than 300bp from each end cannot be sequenced with Illumina technology
- Low complexity libraries pose challenges

Choosing a right approach for your project …. is beyond the scope of this presentation

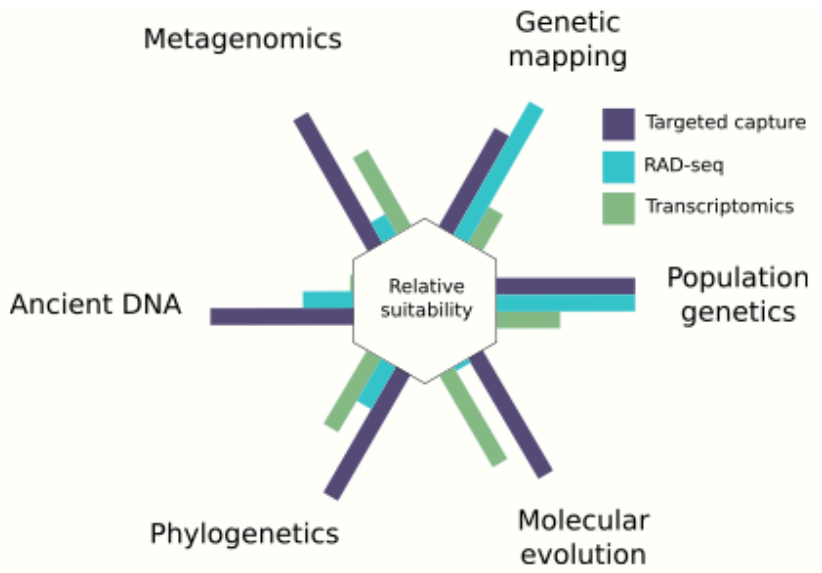


Fig. 4 The relative suitability of targeted capture (purple), RAD-seq (blue) and transcriptomic approaches (green) for genetic mapping of phenotypic traits, population genetics (includes inferring population history and detecting population-level signatures of selection), molecular evolution (e.g. rates of protein evolution), phylogenetics, ancient DNA and metagenomics. Specific height of each bar is arbitrary, but increasing height corresponds to increasing suitability.

Jones & Good, 2016
Targeted capture in evolutionary and ecological genomics

Molecular Ecology
doi: 10.1111/mec.13304

See also:
http://cgrlucb.wikispaces.com/file/view/Applications+of+Sequence+Captures+in+Non-model+Organisms.pdf
especially summary on page 25

How to get from nucleic acid to sequencing

- A library must be first be constructed
- Not an insignificant cost: if multiple samples are run in the same lane (multiplexing), sometimes library preparation costs exceed run costs
- QB3's FGL will prepare libraries starting at $131 each (gDNA) —a good option to consider if you have a large budget and few libraries/little time: http://qb3.berkeley.edu/qb3/fgl/rates.cfm
- Outside providers: rapid-genomics.com (library prep, captures), Cornell (GBS), SNPsaurus (nextRAD), Argonne (amplicon) genohub.com
- Otherwise, purchasing a kit or using the EGL can bring costs done to below $20 each (gDNA), but you have to put the labor in yourself.

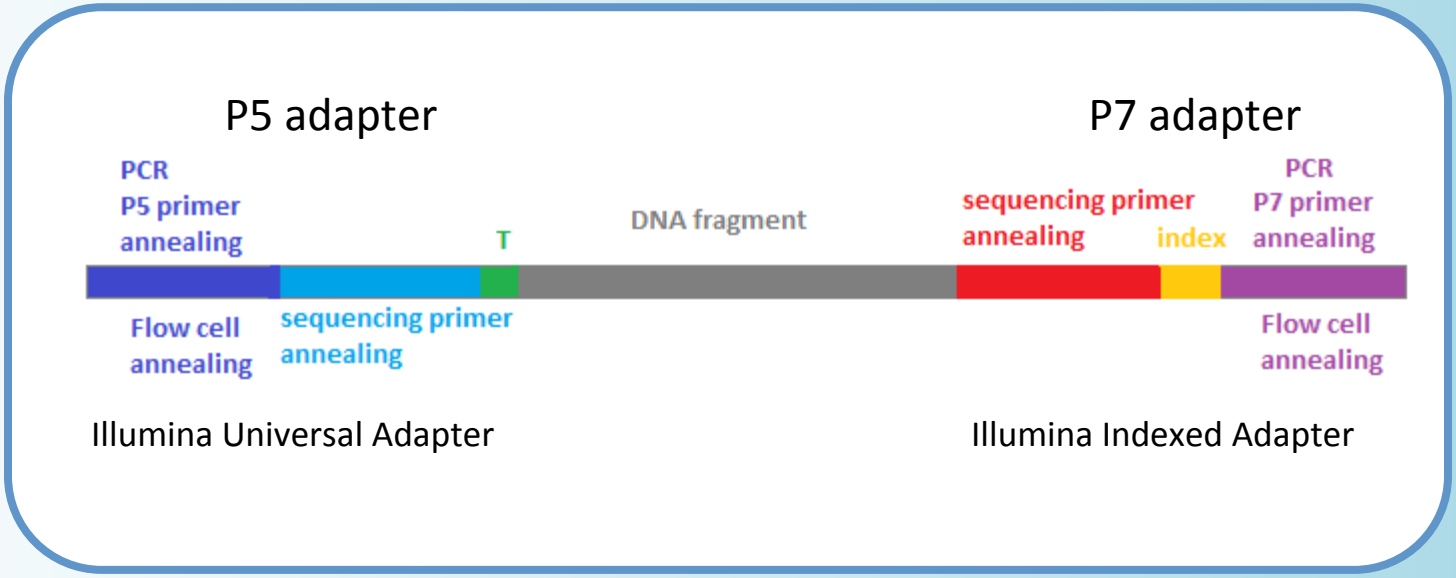
What is a library? (Illumina)

- DNA is fragmented: sonication or enzymatic reaction. (If starting with RNA, it is converted to cDNA before library prep. All Illumina libraries are DNA libraries since they must be double-stranded)
- Ends are repaired to be double-stranded and are prepared for ligation with phosphorylation and/or A-tailing
- Adapters are ligated onto the ends
- PCR enrichment amplifies the number of complete library molecules
- All Illumina library preparation methods can be run on all Illumina platforms (although some may fit better with certain options depending on project design)

Anatomy of an Adapter

An adapter molecule has three parts:

1) flow cell annealing site (outer adapter)
2) index (optional)
3) sequencing primer annealing site (inner adapter)



DNA/RNA Preparation

For most projects, the extraction and assessment of nucleic acids will by far be the most time-consuming part of the lab work.

Sometimes this may feel frustrating—like you are spinning your wheels before getting started on the real work—but DNA/RNA preparation and quality assessment is the most important part of the library preparation process.

Nothing is more correlated with success than starting with sufficient quality and quantity of nucleic acid for your project needs.

The best quality DNA/RNA is going to give us the best sequencing results

DNA Extraction

1) Take care to minimize contamination since every double-stranded molecule inside your sample can be turned into a library.
2) For easier assessment, incorporate an RNaseA digestion step into your library preparation
3) Avoid spin column kits: “death to high molecular weight DNA”
4) Elute/Resuspend your DNA in a buffer with low or no EDTA buffer: 1x LTE (10mM Tris, 0.1mM EDTA) or Qiagen EB (10mM Tris pH 8.5)

DNA Assessment

1) Qubit (don't rely on nanodrop alone): i.e. use dye-based assessment (fluorometer) which will only quantify dsDNA instead of spectrophotometer readings which collect data from any nucleic acid (or anything else absorbing light at 260nM)

The Qubit uses reagents, so there is a charge per use. However, it is worth the money to get accurate **dsDNA** information since that is the only material which will get made into libraries.

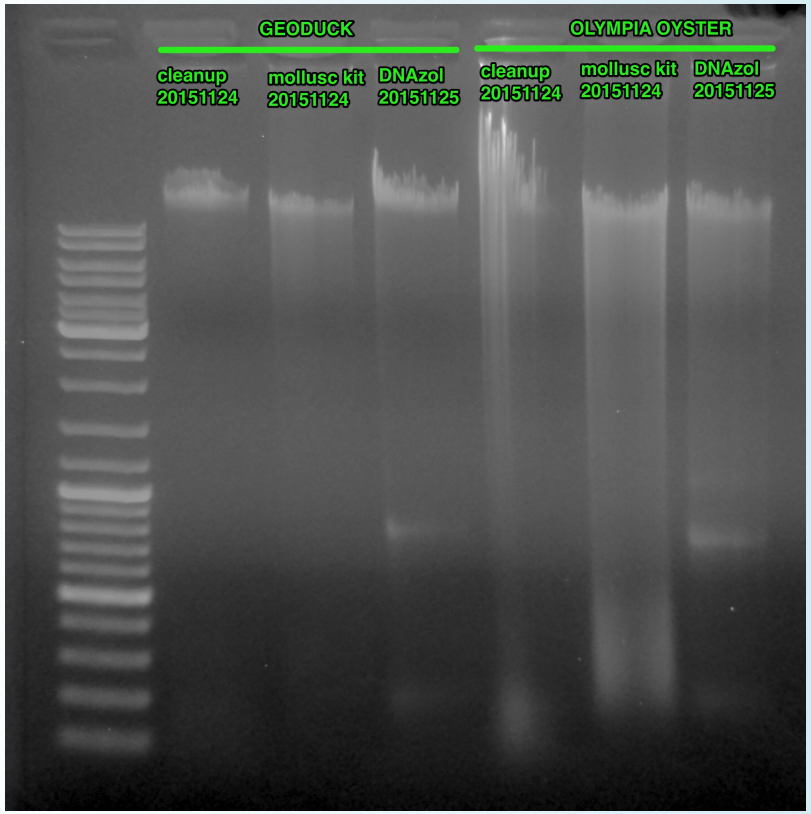
FGL/GSL: $4.00/sample

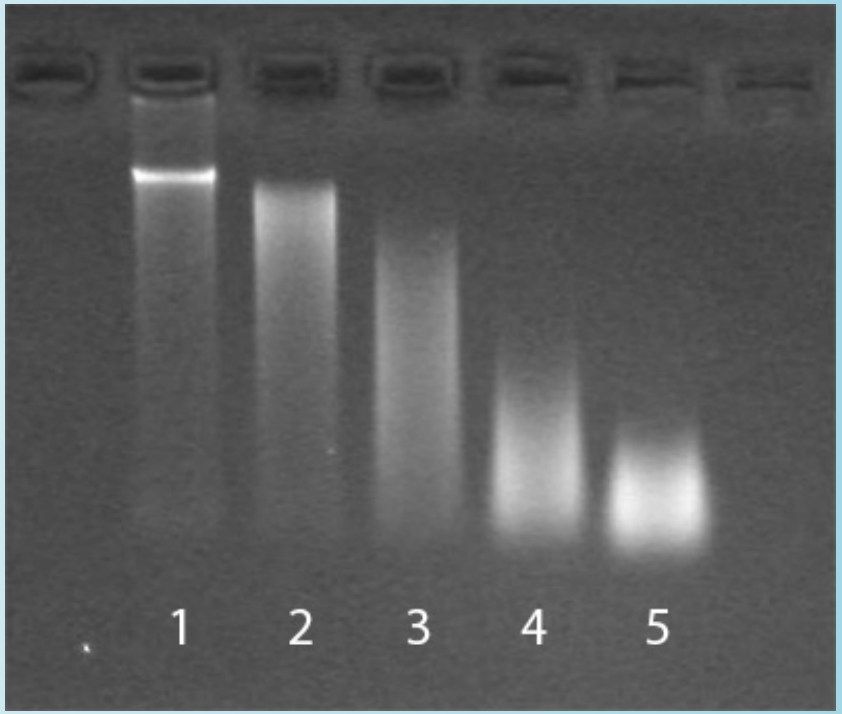
also available in the EGL for members of that lab

DNA Assessment

2) Run DNA samples on an agarose gel to check for high molecular weight DNA and signs of any degraded sample/residual RNA.

Use fresh (non-recycled) running buffer; post-stain in bath if possible





Note: degraded samples <u>can</u> still be used for many library preps, but must be handled differently

DNA Assessment

3) Bioanalyzer (optional for DNA assessment): nanofluidics device that preforms size fractionation and quantification of small samples of DNA, RNA, or proteins.

Will only visualize DNA up to 12kb max, so most DNA is out of range. It is also redundant if you already have sizing data from agarose gel and concentration data from the qubit (considered more reliable).

However, can be useful for getting a more detailed look at the distribution of fragments of degraded samples.

FGL/GSL: $8.00-10.00/sample

also available in the EGL for members of that lab

RNA Extraction

Caution: this is not meant to be an exhaustive list. RNA is far pickier than DNA. If you haven't worked with it before, please consult others with experience for their detailed advice before embarking on collection of tissues meant for RNA research.

1) Proper handling of tissue is key: flash-frozen/RNA Later then stored at -80C or below. Thawed as few times as possible.

2) Be very cautious to avoid environmental RNAses during the extraction process (as well as post-extraction handling): use RNA-restricted equipment & plastics, clean well with Eliminase/RNA-Away, change gloves as often as needed…

3) Homogenize samples well. If possible, use a mixer or bead beater to homogenize rapidly and thoroughly

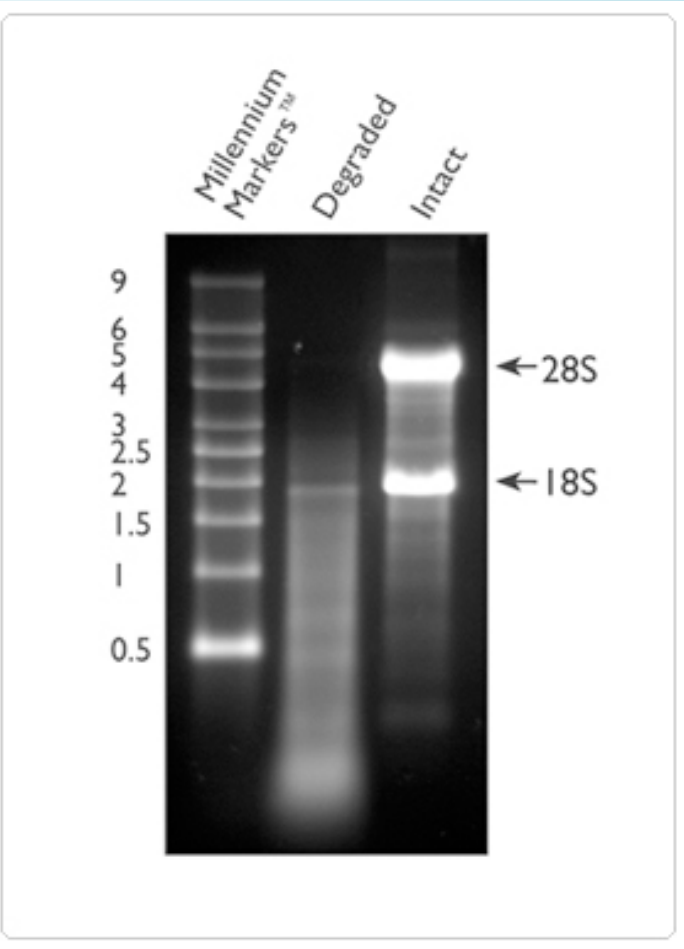
4) Use a Dnase treatment step. (Residual DNA can be turned into libraries)

5) Subsample RNA after extraction so that only a small aliquot must be thawed for QC purposes. Store the main tube at -80C until ready to begin library preparations.

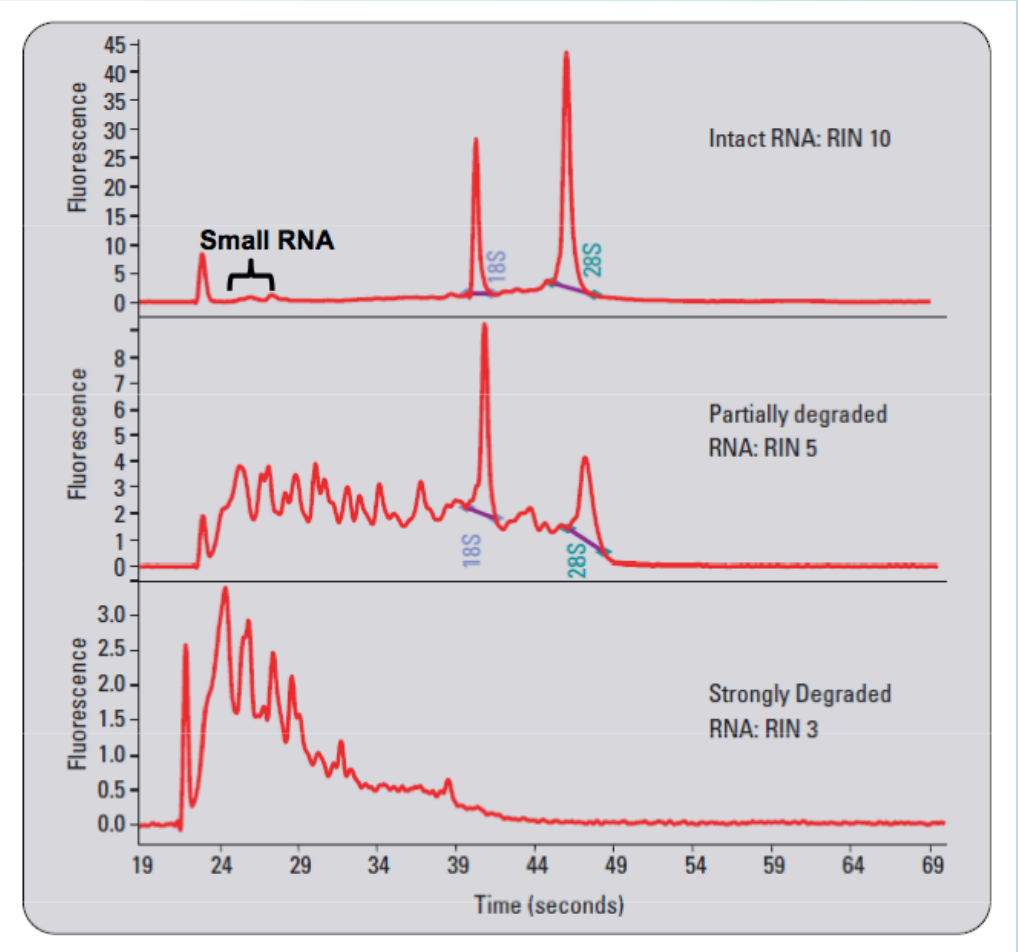
RNALater vs flash-frozen tissue

- RNALater sample handling can be less rushed during RNA extraction. As soon as flash frozen tissue thaws, RNases kick in and start RNA, but RNALater has penetrated the tissue. [Note: this may require some testing to get right and ensure RNALater permeation. For example, some harder tissues may need to be chopped up.]
- RNALater is fieldwork friendly for when LN2 is not available. In some cases samples have gone weeks without refrigeration in warm countries and still produced usable RNA. (That said, freezing is still the best practice for storage after RNA Later has permeated cells.)
- Published "nucleic acid preservation buffer" (doi: 10.1111/1755-0998.12108) *very similar* to RNALater patent ***cough, cough***
- This is not to discourage flash freezing of tissue if that is your protocol but rather to encourage people doing field work to consider tissue collection methods that preserve RNA. It doesn't have to be expensive (see above)

RNA Assessment: agarose gel



RNA Assessment: Bioanalyzer Qualitative and Quantitative



RNA bioanalyzer: quality assessment RNA Integrity Number

- Algorithm designed by Agilent (Bioanalyzer, TapeStation) for their instruments. Base on:
  - rRNA peaks (18S, 28S) are high, indicating that rRNA is still intact
  - how much material is found in the region between the 5s and 18S regions
- Agilent-hosted database to look at user-uploaded, real-world traces: http://www.chem.agilent.com/rin/_rinsearch.aspx
- CGRL consultation
- EGL transcriptome database

RNA bioanalyzer: quality assessment RNA Integrity Number

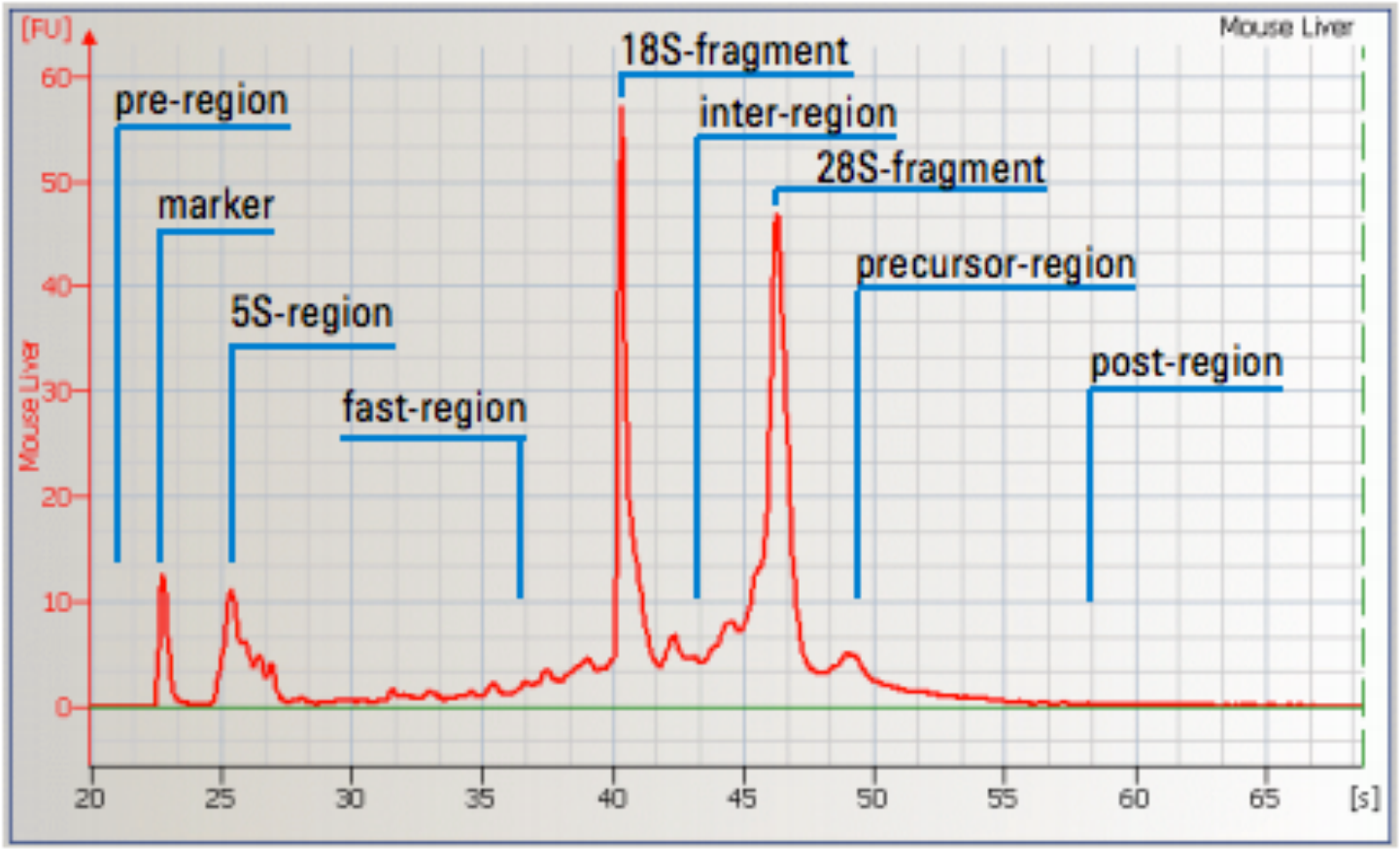
What the RIN can do:

- Obtain an assessment of the integrity of RNA.
- Directly compare RNA samples (e.g. before and after shipment, compare integrity of same tissue across different labs, etc.).
- Ensure repeatability of experiments (e.g. if RIN shows a given value and is suitable for microarray experiments, then the RIN of the same value can *always* be used for microarray experiments given that the same organism/tissue/extraction method was used).

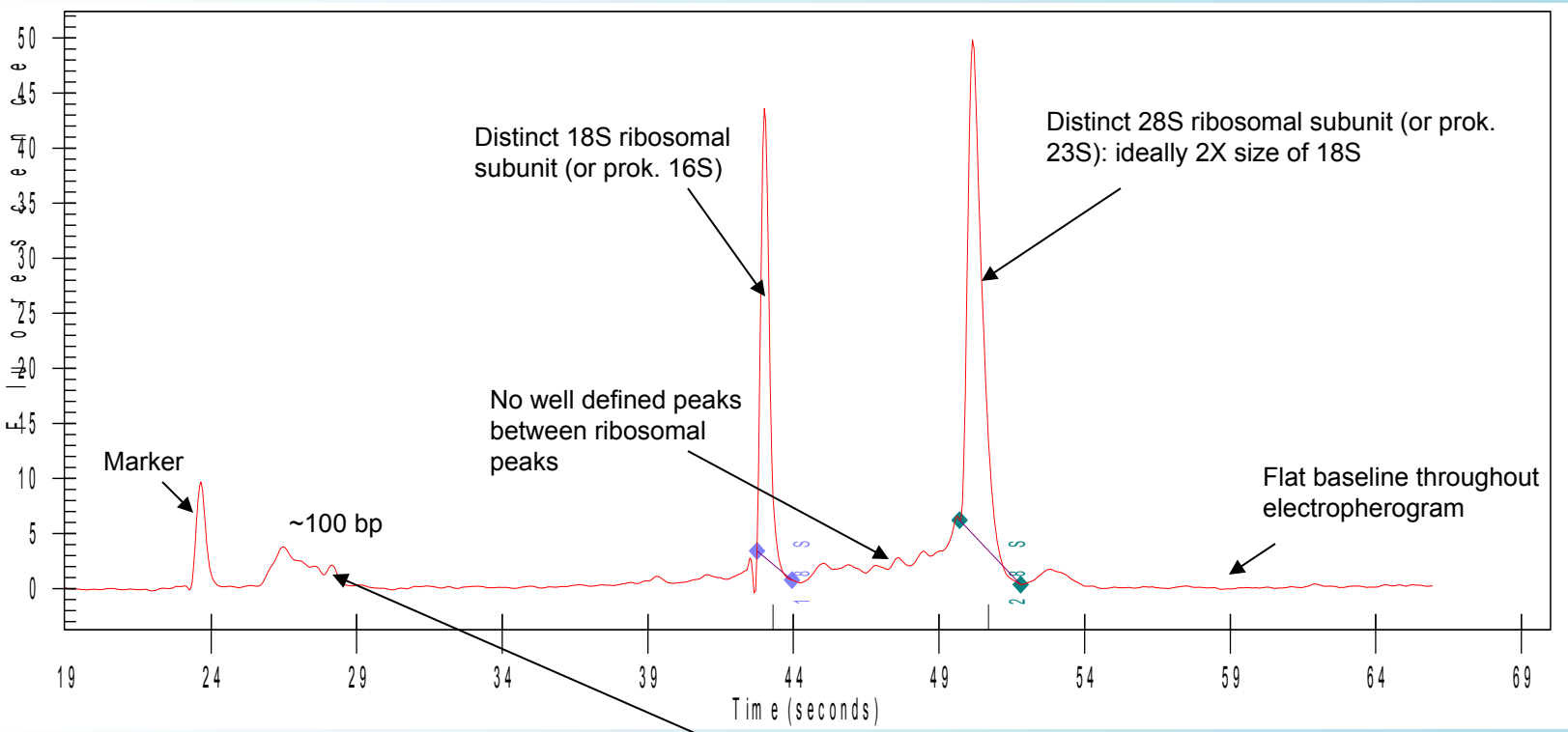
What it CANNOT do:

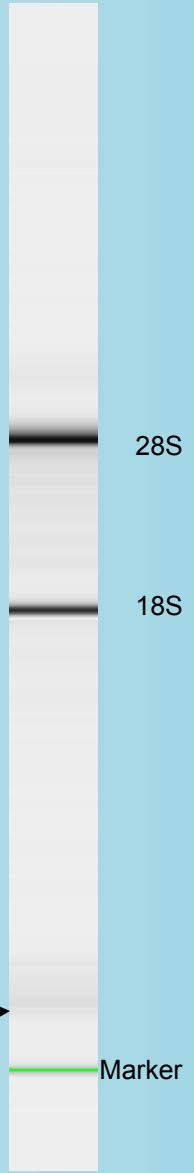
- Tell a scientist ahead of time whether an experiment will work or not if no prior validation was done (e.g. RIN of 5 might not work for microarray experiments, but might work well for an appropriate RT-PCR experiment. Also, an RIN that might be good for a 3' amplification might not work for a 5' amplification).

RNA bioanalyzer: quality assessment
RNA Integrity Number



Intact Total RNA

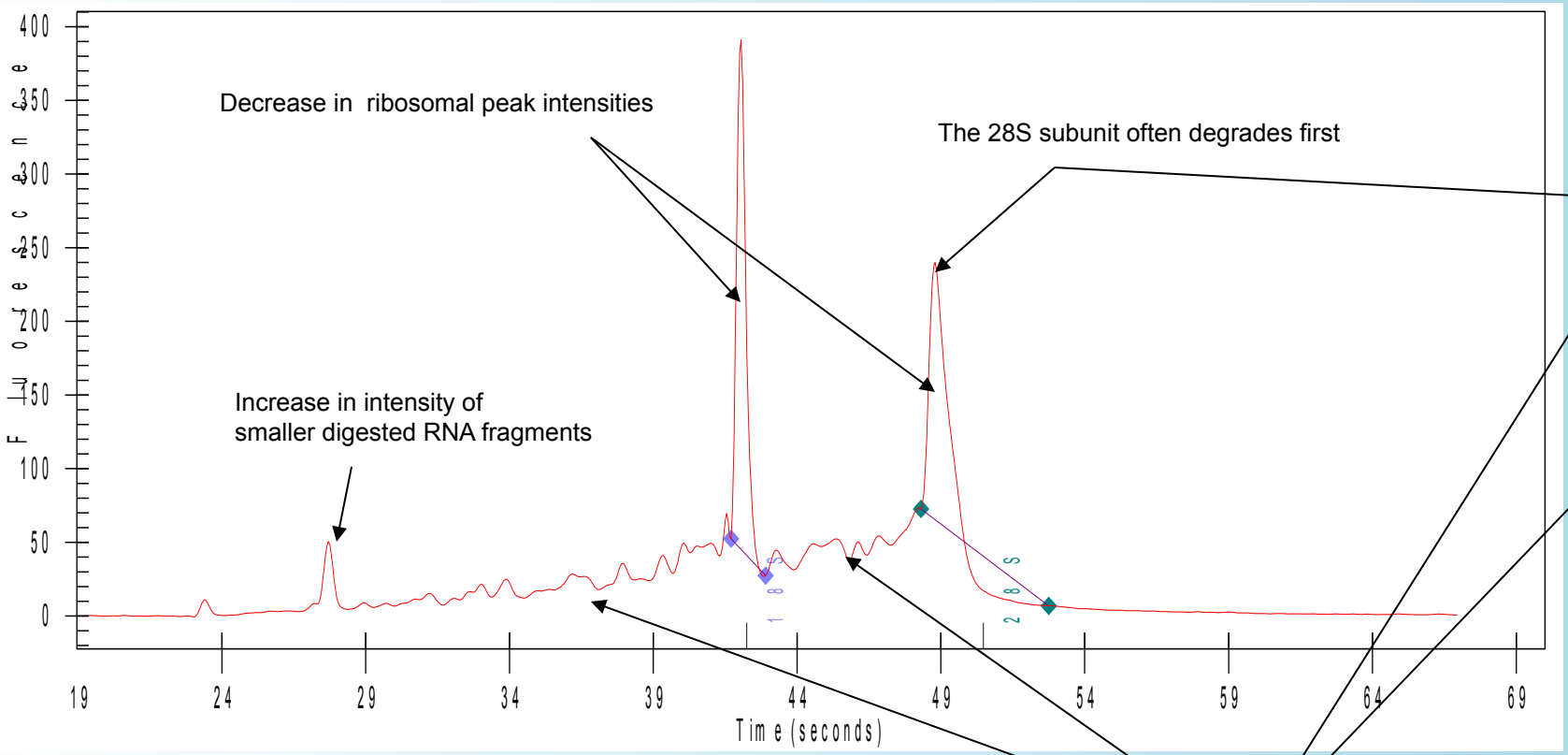




Small peaks are sometimes present after the marker at 24 – 29 seconds. These are represented by 5S and 5.8S subunits, tRNAs, and small RNA fragments about 100bp. These are especially noted when using phenol and trizol extraction methods. They can be removed by treating total RNA through Qiagen columns which removes small RNAs.

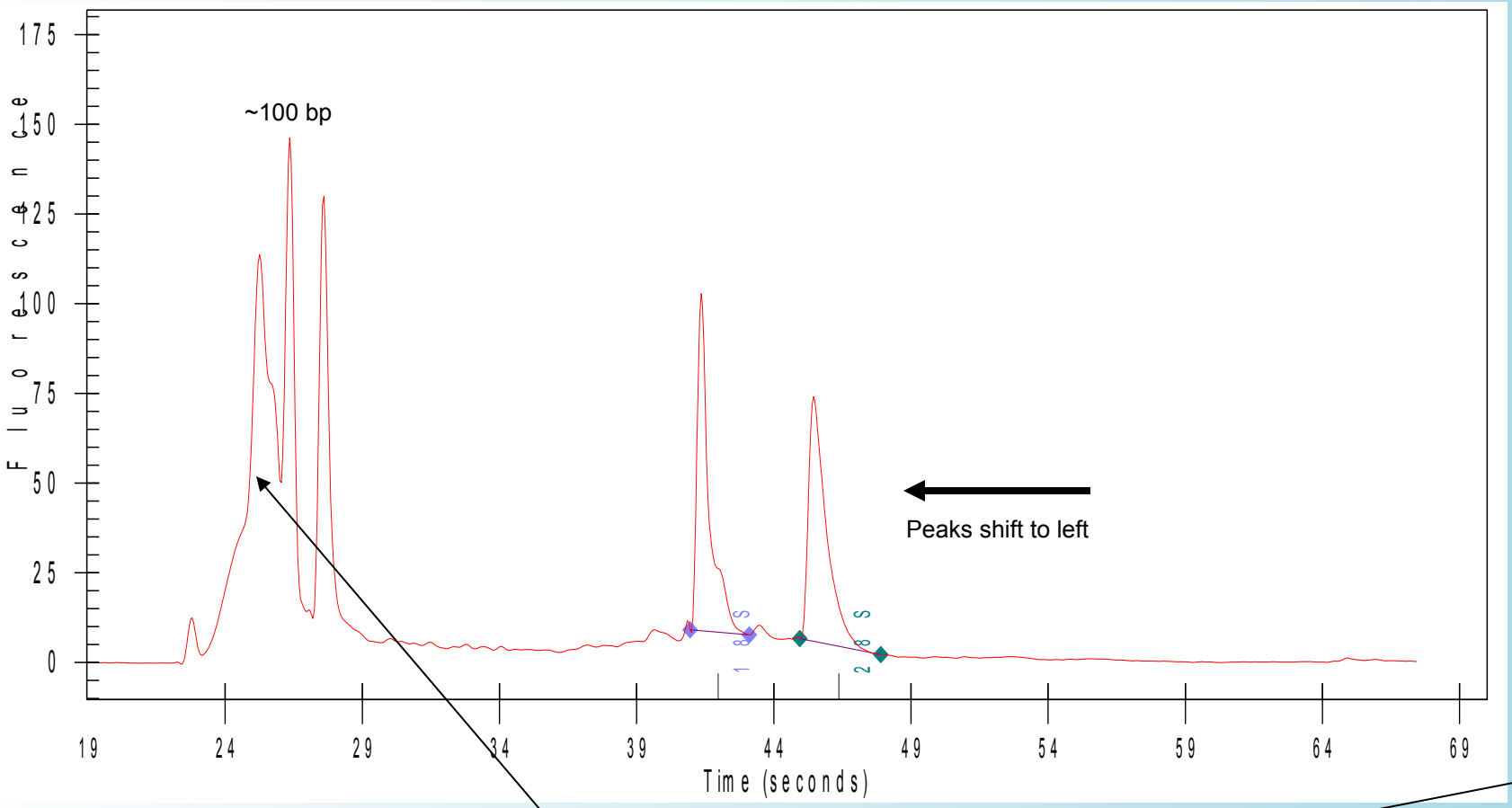
Partially Digested Total RNA

Total RNA with images like this are borderline. Re-extraction should be seriously considered.



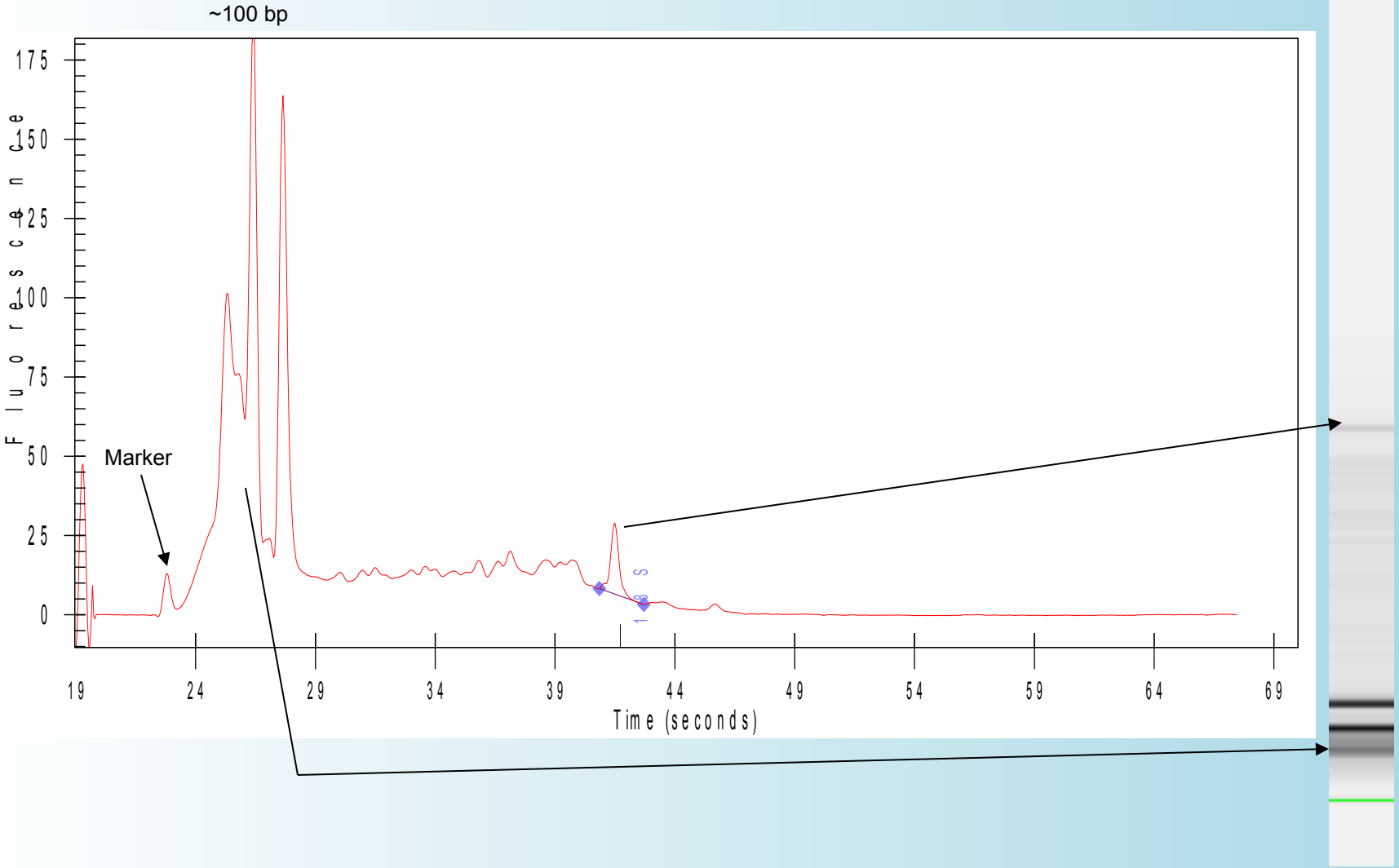
Heavily Digested RNA

Samples of this quality, if labeled and hybed to a chip, might be in question.



High digested RNA peaks

Completely Digested RNA



What if my RIN is poor and my RNA samples appear degraded?

- Substitute other samples, if possible
- Determine risk of proceeding with poly-A selection:
  - looking for rare transcripts or only the most common?
  - using transcriptome data only to sequence exome or for differential gene expression?

If risk is high or if you want to sequence all RNA (viruses, small RNA, non-coding RNA in addition to mRNA), consider a ribozomal RNA depletion method

OK, I have my high quality DNA/RNA extracted and assessed in pain-staking detail, now can I start making a library?

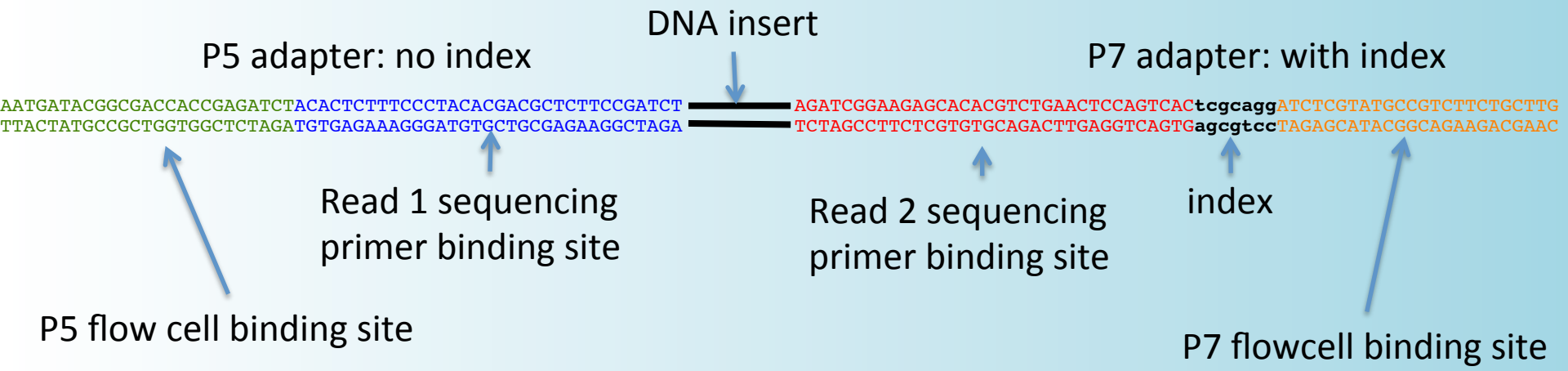
Yes, but only if you are sure of your:

1) sequencing platform & number of lanes
2) library preparation method
3) number of samples
4) indexing plan
5) amount of coverage
6) etc, etc, etc.

What is an adapter? (Illumina)

An adapter molecule has three parts:

1) flow cell binding site (outer adapter)
2) index (optional, but recommended)
3) sequencing primer binding site (inner adapter)



What does the index do?

In Illumina sequencing, an index is a short (6-8bp) unique tag of DNA incorporated into the adapter

The index sequence is read separately from the library insert, but the tag information is linked in the metadata of each sequencing read

After a run containing multiple libraries, all the data is demultiplexed: all sequencing reads with the same index tag are put into the same folder

This process allows more than one library to be sequenced in a single Illumina lane: 2, 60, 100's …

Single vs Dual Index

In order to save money on index costs, some protocols use a dual index approach. If there are 12 unique P7 indexes and 8 unique P5 indexes, then 96 combinations exist for the cost of only 20 oligos

With either approach, be meticulous about your recordkeeping. The index is the only link between your sequencing data and its biological meaning

Index incorporation

Indexes are incorporated into the library adapter in two ways as you will see in the section following:

1) The index is in an external oligo which extends an initially short adapter to be full-length (moderately expensive ~$40 each)
2) The index is already incorporated in a long adapter complex (more expensive: ~$120 each)

For projects that need a lot of indexes, it isn't usually cost-effective to purchase them individually. Instead, some library prep kits come with indexes (often 12 single or 96 dual) and indexes can often be purchased from larger labs that have a greater volume of library preps (some options are noted in the following slides)

Selecting indexes

1) Ideally, each library in a project should have a unique index or combination of indexes (if using a dual indexing strategy) even if you plan to run samples over multiple lanes. That way, if plans change, you can be flexible about how samples are combined.

2) If planning a capture experiment, the indexes you use must be compatible with available blocking oligos

3) Be careful to keep meticulous records with respect to which index gets incorporated into which library.

Balancing indexes 1

Index combinations should:

1. Have sequences as distinct from each other as possible
2. Have a relatively even representation of all 4 bases at each position
3. Must have A/C and G/T in each position of the index

Table 4: Examples of Good and Bad Index Combinations

Good Examples				Bad Examples			
Index 1		Index 2		Index 1		Index 2	
705	GGACTCCT	503	TATCCTCT	705	GGACTCCT	502	CTCTCTAT
706	TAGGCATG	503	TATCCTCT	706	TAGGCATG	502	CTCTCTAT
701	TAAGGCGA	504	AGAGTAGA	701	TAAGGCGA	503	TATCCTCT
702	CGTACTAG	504	AGAGTAGA	702	CGTACTAG	503	TATCCTCT
	✓✓✓✓✓✓✓✓		✓✓✓✓✓✓✓✓		✓✓✓✓✓✓✓✓		✓✓✓✓XXXX

✓= signal in both color

X= signal missing in one color channel

Balancing indexes 2

This is especially important when using just a few (2-8) indexes at either the P7 or P5 position because the odds are higher that by chance you will pick a poor combination

You can usually find lowplex pooling information from the company that provides the indexes or the protocol that they were synthesized from.

Or check with the published protocol or lab you purchased indexes from. For Meyer & Kircher indexing oligos used in the EGL, if used in order starting with indexing1, they will be balanced

Index vs Internal Barcode

- Index is read separately from sequencing data and is bioinformatically linked
- Internal barcode is read by the sequencer at the start of the read.
- A small amount of read length is lost with an internal barcode, but that is off-set in projects with low complexity libraries by advantages:
  - Combinatorial coding allows unique combinations of internal barcode + external adapter (samples can be pooled earlier in the library prep process to save money)
  - Internal barcode creates sequence diversity at crucial initial bases for RAD-Seq and amplicon libraries (most other types have naturally high complexity)

How to incorporate the adapter and index?

Depends on the type of data to be collected:

- genomic DNA (whole genome sequencing, targeted capture)
- RNA sequencing
- RAD-Seq
- Amplicon sequencing
- ChIP-Seq
- Mate Pair

Illumina library preparation: gDNA
1) Two-stage library preparation

Meyer & Kircher 2010: doi:10.1101/pdb.prot5448

iTru/Adapterama system:
https://conf.abrf.org/sites/default/files/images/tglenn_abrf_adapterama_march_2015_final.pdf

NEBNext

Pros:

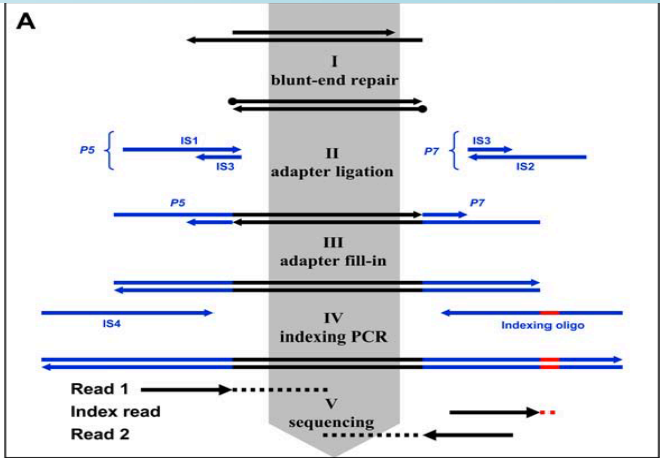
- No kit required (reagent costs in EGL = $10-15 per sample) or use of inexpensive Kapa Biosystems kits that do not come with adapters
- Multiplexing: the sky's the limit! Single internal adapter used in combination with many index combinations. These are expensive to buy in large quantities but can be purchased from other lab groups:
  - EGL (Meyer & Kircher protocol): 105 P7 indexes; 8 P5
  - GSL: will sell adapter stubs and plates of 96 indexing oligos; e-mail Shana for details
  - Glenn (iTru): 48 P7; 48 P5 (http://baddna.uga.edu/)
- Degraded DNA, historical samples work just fine (sometimes better)
- Samples can be used for downstream enrichment by targeted capture

Cons:

- Best practice is 1000ng of starting material if libraries will be captured(usually easy to come by but not in all cases). Less can be used if not available for WGS.
- Has more steps than the other gDNA methods, so more time in the lab
- Requires use of a sonicator or fragmentase
- PCR step is required to extend the adapter and incorporate the index (only a con if a no-amplification method is necessary)

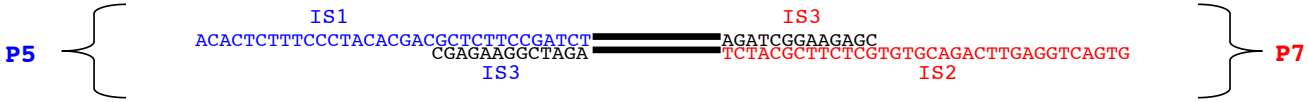
Illumina library preparation: gDNA
1) Two-stage library preparation

Meyer & Kircher protocol (2010)



1) stubs including the internal adapter sequence are ligated

Adapter ligation:



Adapter fill-in:

Indexing PCR:

2) adapters are extended to full length and an index is incorporated with PCR

Library with adapters:

Illumina library preparation: gDNA
2) Y-shaped adapter

Illumina TruSeq kit

Other vendor kits (Kapa, NEB) can be used but adapters much be purchased separately (dual index Y-shaped adapters will be sold by GSL in plates of 96. E-mail Shana for details)

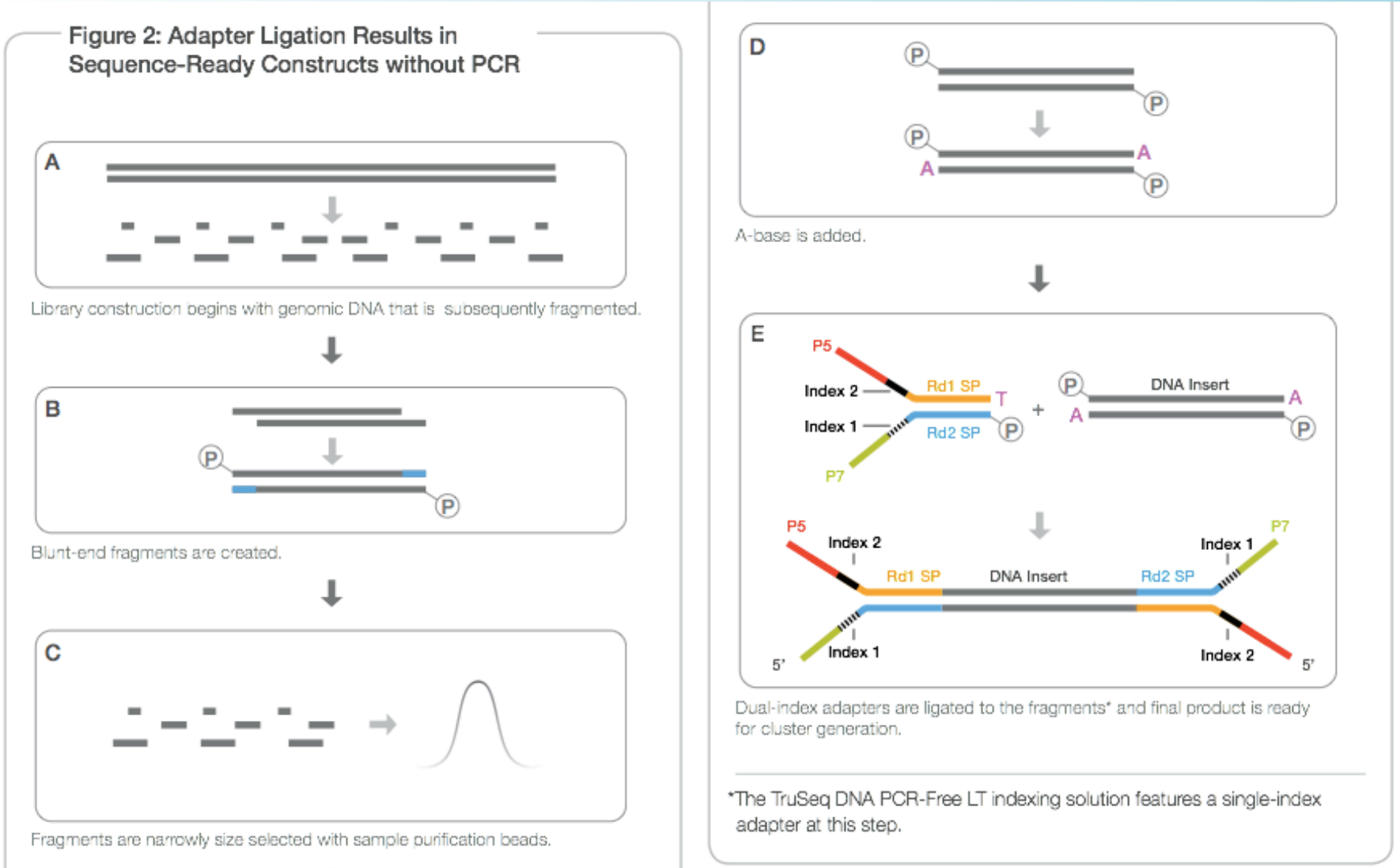
Pros:

- Slightly fewer steps than the two-stage library prep; higher yields
- PCR enrichment step is not necessary since the library is full length after adapter ligation; can be used for library preps when PCR should be avoided or minimized so as not to introduce biases
- Best choice for WGS, especially with large genomes. An amplification-free approach eliminates PCR biases caused by difficult to amplify genomic regions
- Degraded DNA, historical samples work just fine
- Samples can be used for downstream enrichment by targeted capture

Cons:

- Requires 500-1000ng of starting material (usually easy to come by but not in all cases)
- Adapters are expensive since each one must be individually barcoded.
- (however, GSL may start selling aliquots of Y-shaped adapters
- Requires use of a sonicator or fragmentase

Illumina library preparation: gDNA
2) Y-shaped adapter



Illumina library preparation: gDNA
3) Nextera

Illumina Nextera: reagent and index costs ~$100/sample (smallest kit: 24)

Pros:

- Least number of hands on steps in the lab→ very fast (as low as 90 minutes of lab time)
- Does not require a sonicator or fragmentase
- Protocol requires very small amounts of starting material

Cons:

- Cannot be used with degraded/historical DNA
- Not recommended for downstream targeted capture due to low starting material (except for human Nextera Rapid Capture Custom Enrichment Kit)
- PCR step is required (only a con if a no-amplification method is necessary)
- May displace some biases in transposon cut sites
- Most expensive method, only sold by Illumina. However, some users have found that they can split one preparation into 5 or even 10 samples to make the per sample price quite low.

Illumina library preparation: gDNA
3) Nextera

Transposomes:
1) Fragment
2) End Repair
3) Ligate Internal Adapters

All in one step!

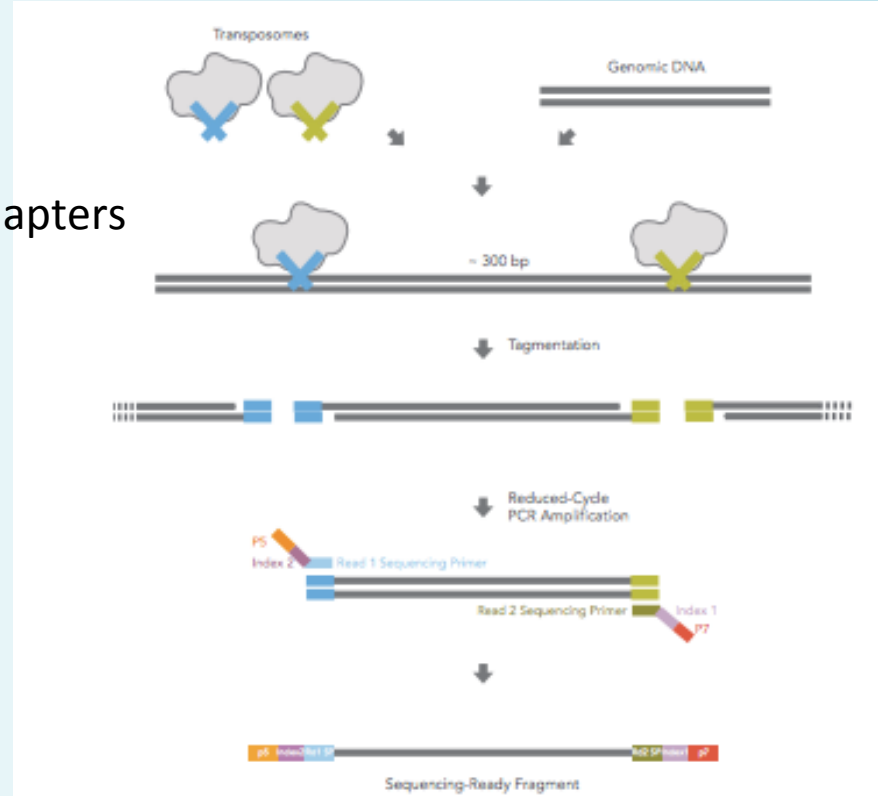


Figure 2: Nextera Library Preparation Biochemistry—Nextera chemistry simultaneously fragments and tags DNA in a single step. A simple PCR amplification then appends sequencing adapters and sample indexes to each fragment.

Then the index and external adapter is incorporated with PCR

DNA fragmentation: gDNA Two-stage library preparation & Y-shaped adapter

Sonication:

- Covaris is available in the FGL
- Bioanalyzer is available in the EGL

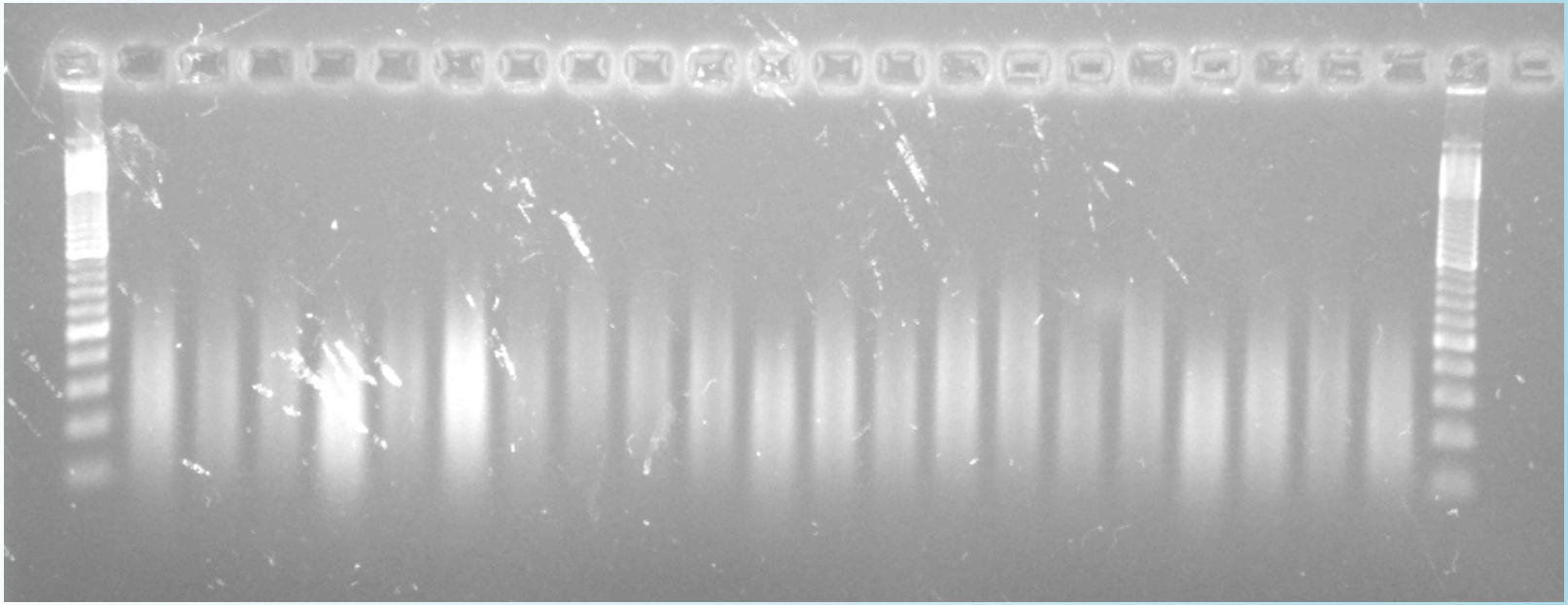
Fragmentase:

- Sold by New England Biolabs
- Part of Kapa Biosystems Hyper Prep Kit

Both enzymatic and sonication methods may require a fair amount of optimization to figure out the right conditions. However, if all samples are handled the same way (concentration, dilution buffer, etc), hopefully those conditions, once determined, can be reliably used for most samples of a project.

DNA fragmentation: gDNA

The ideal DNA smear distribution will depend on the type of data being collected, the length of the sequencing run planned, and whether you want PE reads to overlap or prefer that they don't (wasted data)



Post-sonication gel image for planned exon capture experiment to be sequenced with PE100. Aim is for an average insert size around 250 so that most reads do not overlap in the middle but so that insert sizes stay relatively small (smaller libraries capture better and can be run on the higher throughput instrument HiSeq4000).

Museum/Historical Samples

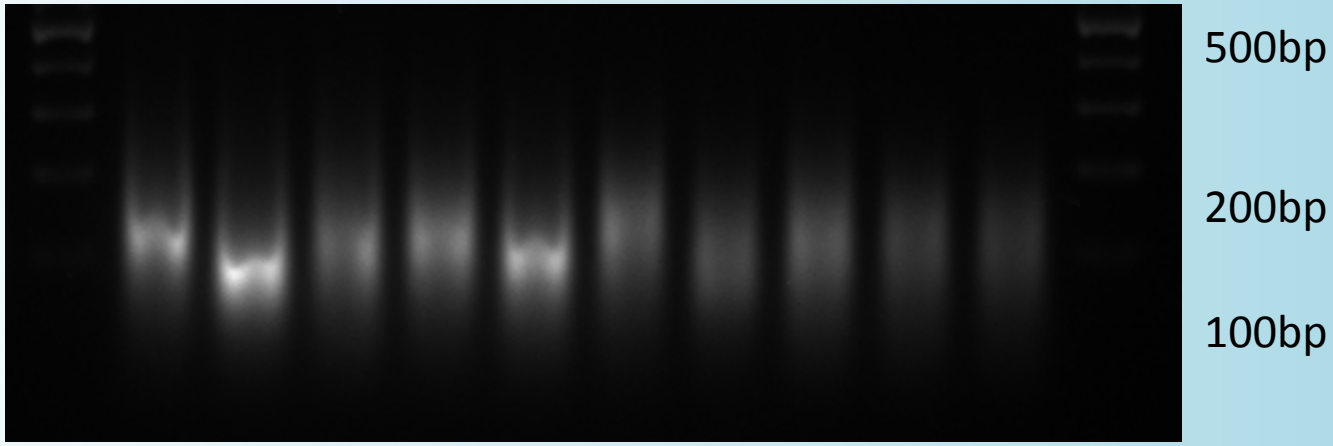
DNA extracted from tissue sources not collected or preserved for that purpose (study skins, hair, toe pads, formalin-fixed specimens, pinned insects/spiders, herbaria specimens) but for other scientific study in museums

Historical DNA is a subset of ancient DNA (DNA recovered from biological samples that have not been preserved specifically for later DNA analyses)

Older ancient DNA samples have been exposed to the environment longer after death and require more specialized lab facilities for safe handling (positive airflow, UV lamps, full body PPE.)

Moderate quantities of DNA can be extracted from museum samples of the past ~100 years: need for an isolated space, separate reagents, and extreme caution when handling samples.

Historical/Degraded DNA



Degraded samples may still be appropriately sized for short-read sequencing

Sometimes they do not even need sonication!

Must be quantified with HS qubit (or other fluorometer)

Ideal for targeted capture experiments; if more than partially degraded, not a good fit for RAD-Seq unless combined with a capture approach (HyRAD, Rapture); impossible for RNA-Seq

Historical DNA: challenges

Destructive sampling: often museums wary of “loaning” materials without proof-of-concept for the work

Sometimes samples are **very** fragmentedand produce **very** low yields. Difficult to trouble-shoot/optimize

Higher likelihood of environmental contamination

Caution not to lose valuable fragments during SPRI bead clean-ups (increase ratio as needed)

Some preservation methods introduce contaminants and complications beyond fragmentation of DNA

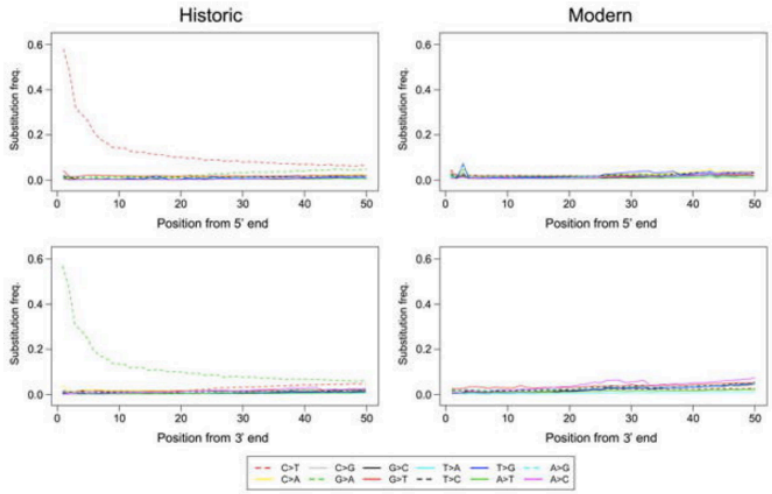
- Ex. Formalin-fixed materials extra challenging due to cross-linking with protein in addition to degradation.
- DOI: 10.1371/journal.pone.0141579

Higher amounts of adapter contamination in sequences (can be bioinformatically removed, but loss of data)

Historical DNA: challenges

Post-mortem DNA damage caused by hydrolysis often converts cytosine to uracil. Subsequent polymerases replace uracil with thymines giving the appearance of a spurious C → T substitution (or G → A in reverse strand)

1) Use proofreading polymerases that stall in the presence of uracil to reduce (but not eliminate) misincorporation errors [note: you should use proofreading polymerase for all library preparations involving PCR]
2) These substitutions can be filtered from the analysis
3) Ends of reads can be trimmed if samples were not sonicated (but that may mean a significant % data loss)



Patterns of mismatches in historic and modern *Tamias alpinus* sequences

The frequencies of the 12 types of mismatches (*y* axis) are plotted as a function of distance from the 5′ and 3′-ends of the sequence reads (*x* axis). The first 50bp of the sequence reads are shown. The frequency of each particular mismatch type is calculated as the proportion of a particular alternative (non-reference) base type at a given site along the read, and is coded in different colors and line patterns explained at the bottom of the plots: "X > Y" indicates a change from reference base type X to alternative base type Y.

Bi, et al. 2013
Unlocking the vault:
next generation
population genomics

doi: 10.1111/mec.12516

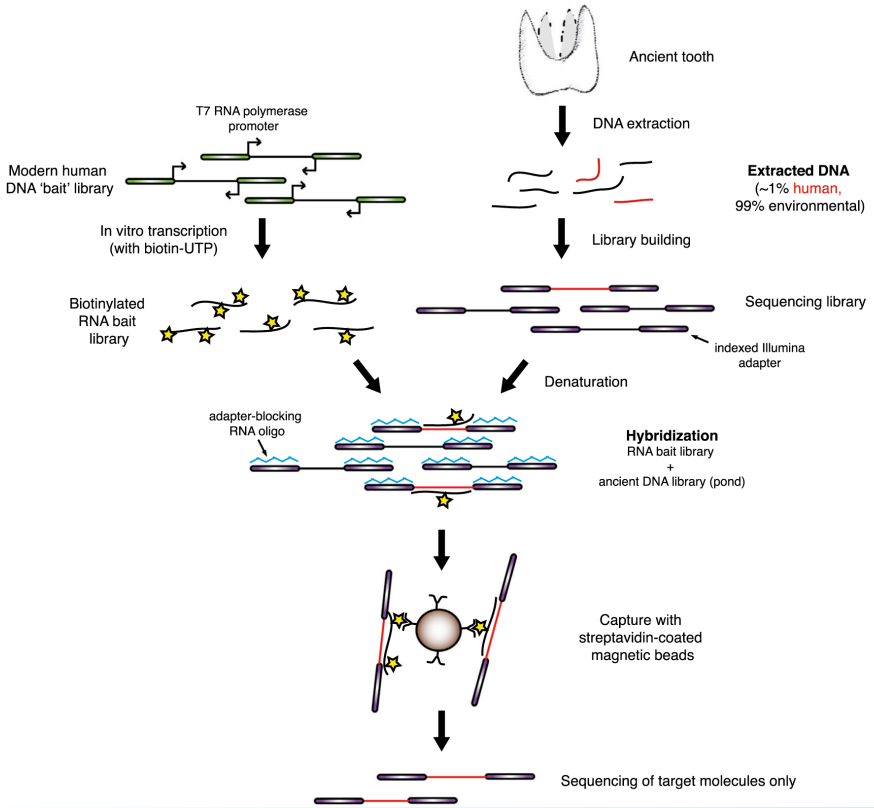
Target Enrichment - Whole Genome In-Solution Capture

Please cite this article in press as: Carpenter et al., Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries, The American Journal of Human Genetics (2013), http://dx.doi.org/10.1016/j.ajhg.2013.10.002

ARTICLE

Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries

Meredith L. Carpenter,[1] Jason D. Buenrostro,[1,14] Cristina Valdiosera,[2,3,14] Hannes Schroeder,[2] Morten E. Allentoft,[2] Martin Sikora,[1] Morten Rasmussen,[2] Simon Gravel,[4] Sonia Guillén,[5] Georgi Nekhrizov,[6] Krasimir Leshtakov,[7] Diana Dimitrova,[6] Nikola Theodossiev,[7] Davide Pettener,[8] Donata Luiselli,[8] Karla Sandoval,[1] Andrés Moreno-Estrada,[1] Yingrui Li,[9] Jun Wang,[9,10,11,12] M. Thomas P. Gilbert,[2,13] Eske Willerslev,[2,15] William J. Greenleaf,[1,15,*] and Carlos D. Bustamante[1,15,*]



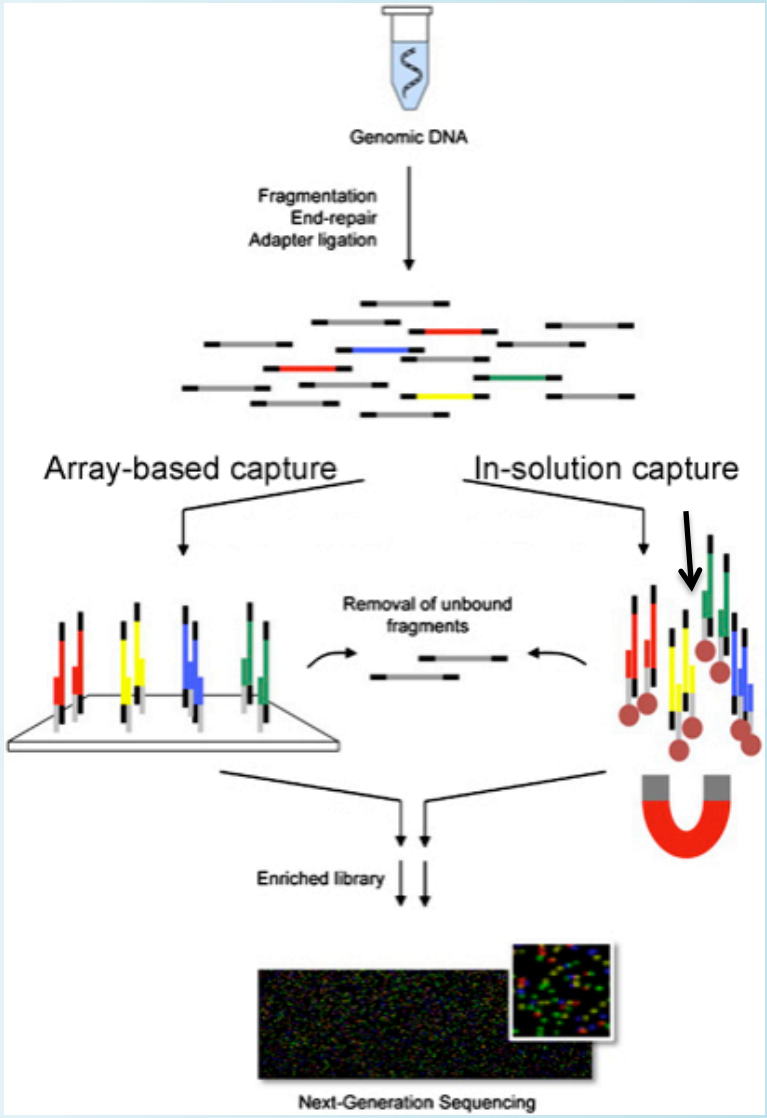
WISC works effectively for enriching genomic DNA from ancient specimens that contain very low levels of endogenous DNA (<1%)

Can increase unique reads 2- to 10-fold

In-house preparation from modern genomic DNA or commercial probes (Mybaits)

doi: 10.1016/j.ajhg.2013.10.002

Hybridization-based target enrichment



60bp unlabeled DNA probes

55-180bp biotinlylated RNA/ DNA probes

DOI: 10.1111/mec.13304

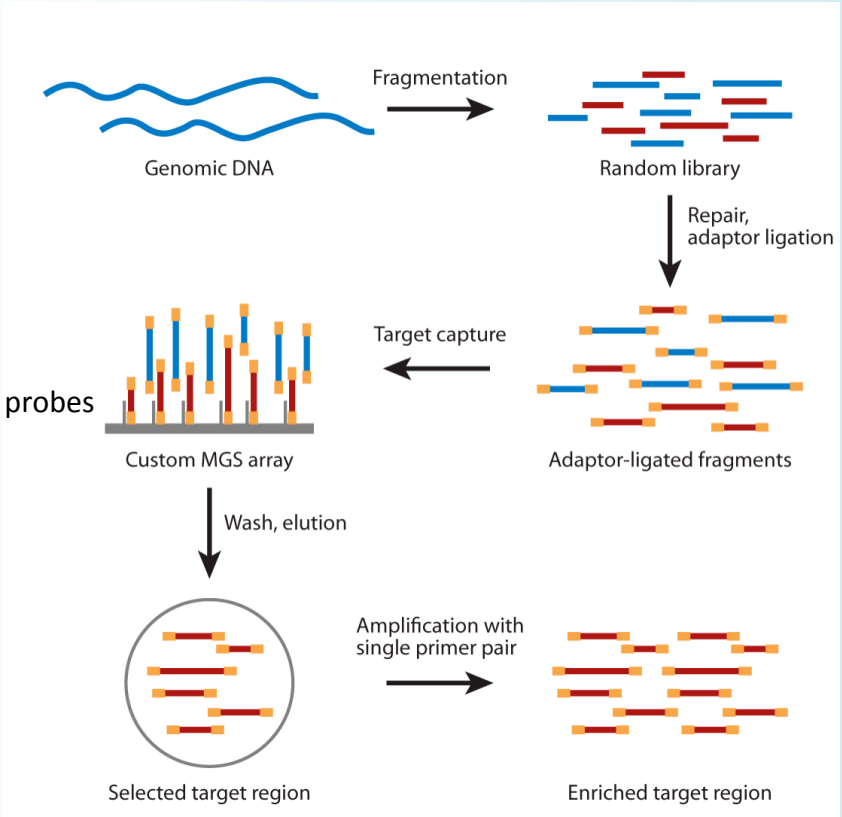
Baits synthesized by a commercial vendors (ie. Roche (Nimblegen), Agilent, Illumina, IDT, MYcroarray)

Target Enrichment – Microarray-based Capture

Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing

Emily Hodges[1,2], Michelle Rooks[1,2], Zhenyu Xuan[1], Arindam Bhattacharjee[3], D Benjamin Gordon[3], Leonardo Brizuela[3], W Richard McCombie[1] & Gregory J Hannon[1,2]

Nature Protocol 2009 4:960-974.



slide courtesy of Ke Bi

Agilent Custom SureSelect microarray 1M or 244K format

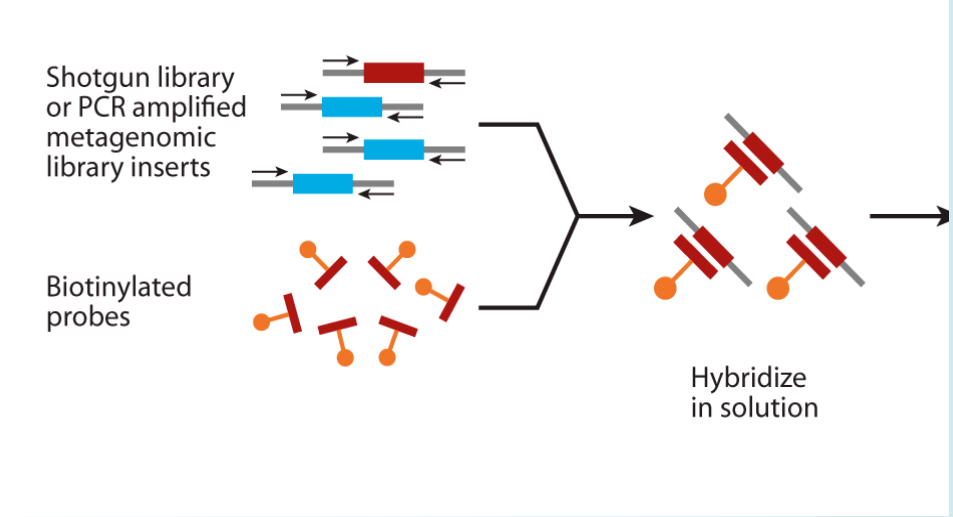
Pros

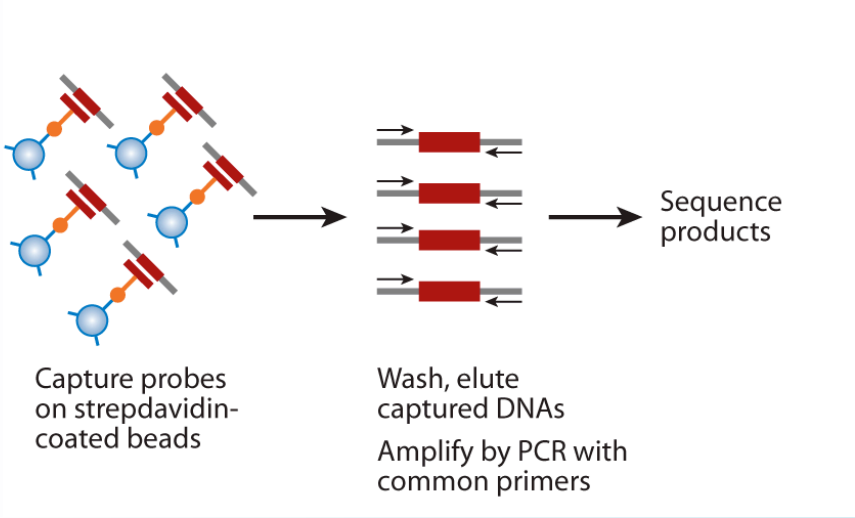
- Low cost: ~750USD for each 1M-probe array
- Suitable for small-scale phylogenomic and population genomic studies
- High probe tiling density/direct control over probe design

Cons

- Need a reference for probe design (also true for other types of hybridization-based methods.)
- Low capture efficiency
- Probe length short (60bp)
- Need special equipments (hybridization chamber, gasket slides, oven, etc. Available in EGL)
- Not cost-effective for surveying large number of samples
- Need large amount of input DNA (20ug/array) and Cot-1 DNA (50ug/ul)
- Complicated workflow

Target Enrichment - In-solution Capture





For non-model systems:

- NimbleGen SeqCap EZ Developer kits
- Mycroarray MyBaits kits

Pros

- Target size large (up to 50 Mb for NimbleGen) or medium but with 10s of captues (MyBaits)
- Low amount of input DNA and Cot-1 DNA
- High level of multiplexing (Nimblegen >50)
- Suitable for large-scale population genomic projects (NimbleGen) or phylogenomic projects (MyBait)
- High capture efficiency
- No special equipment needed)

Cons

- High initial investment (kits are more expensive than single array capture)
- Not cost-effective for multiple, small-scale population genomic projects when distinct target sets/design is required

slide courtesy of Ke Bi

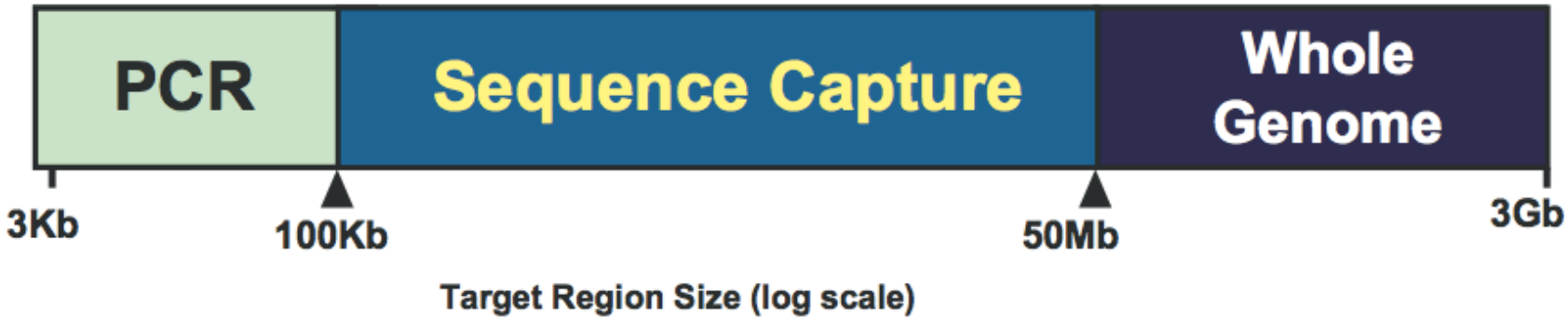
Nimblegen (Roche) EZ Developer

Smallest custom kit size:
4 reactions $7-8000 (usually 5 reactions possible)
12 reactions $12,000 (14-15 reactions possible)

2.1 million unique probes in solution (100's or 1000's of copies of each probe)
Designed in consultation with Roche bioinformaticians from fasta files
Multiplexing of large numbers of samples (> 50) due to high amount of tiling density possible with that number of unique probes
Pre-designed kits available for some model organisms

Custom

- Custom sequence capture allows researchers to specify their own target regions of the genome.
- Researchers often use custom sequence capture as follow-up to Exome resequencing or CGH/SNP array studies.
- The opportunity for custom sequence capture spans from 100kb to 50Mb.
- Analyzing small target regions reduces resequencing costs even compared to exome sequence capture.



Mycroarray MyBaits Custom Kits

Custom kit options start at $2400 (some discounts available)

Number of reactions	Baitset Tier *(Maximum # of bait sequences)*					
	MYbaits-1 *(20,000)*	MYbaits-2 *(40,000)*	MYbaits-3 *(60,000)*	MYbaits-4 *(80,000)*	MYbaits-5 *(100,000)*	MYbaits-10 *(200,000)*
12	$2,400	$3,000	$3,600	$4,200	$4,800	$7,200
24	$3,600	$4,500	$5,400	$6,300	$7,200	$10,800
36	$4,740	$5,924	$7,109	$8,294	$9,478	$14,218
48	$5,760	$7,200	$8,640	$10,080	$11,520	$17,280
96	$8,640	$10,800	$12,960	$15,120	$17,280	$25,920
192	$13,440	$16,800	$20,160	$23,040	$26,880	$40,320
384	$23,040	$28,800	$34,560	$38,400	$46,080	$69,120
768	$38,400	$48,000	$57,600	$65,280	$76,800	$115,200

- In-solution probes in modules of 20,000 designed from fasta files of targets
- Multiplexing is possible, but in smaller numbers (< 10) due to fewer unique probes
- Very cost-effective for many captures (price per capture decreases as more reactions are ordered): great choice for phylogenetics since more closely related species can be captured together, reducing the risk of one library out-competing others

http://www.mycroarray.com/mybaits/mybaits-planning-your-project.html
http://www.mycroarray.com/mybaits/mybaits-custom-calculator.html

Predesigned Kits

If your project design allows, ordering pre-designed kits is the least expensive way to undertake an in-solution targeted enrichment project.

- Nimblegen, Illumina, Agilent: human exon, UTR + exon,
- Agilent, Illumina: mouse exon
- MyBaits: Whole Genome Enrichment
- MyBaits: Mitochondrial bait (planned for chicken, many mammals)
- MyBaits: Ultaconserved Elements
  - Tetrapods
  - Ray-finned fish
  - Hymeoptera

Ultraconserved Elements (UCEs)

- Highly conserved regions of organismal genomes shared among evolutionary distant taxa
- UCE probes are anchors to capture adjacent genomic areas which are more variable (unlike other capture methods, we are interested in the flanking regions here)
- Developed for resolving deep phylogenies but flanking sites may have enough variation for population genetic/ phylogeographic analyses as well. (Leaché et al. 2015: Study using/comparing both RAD-markers and UCEs. doi: 10.1093/gbe/evv026)
- Great website (ultraconserved.org) with protocols, papers/ talks, FAQs, and probe sequences
- Because probes are published, researcher can custom order a subset (as with Leaché et al), include them with other capture probes, or order a pre-made kit from MyBaits

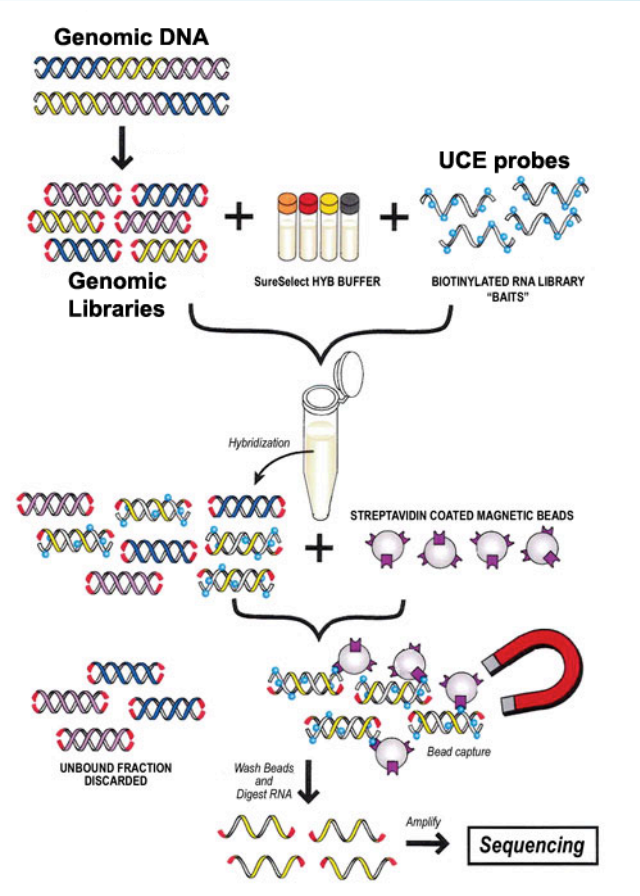
Target Enrichment - Ultraconserved Elements (UCEs) Capture

Syst. Biol. 61(5):717–726, 2012

DOI:10.1093/sysbio/sys004
Advance Access publication on January 9, 2012

Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales

BRANT C. FAIRCLOTH[1,*], JOHN E. MCCORMACK[2], NICHOLAS G. CRAWFORD[3], MICHAEL G. HARVEY[2,4], ROBB T. BRUMFIELD[2,4], AND TRAVIS C. GLENN[5]



- Mycroarray MYbaits-UCEs kits
- RapidGenomics (outsourcing your samples)

Pros

- Cheap (e.g. 5K loci kits cost about $700).
- No need for marker selection and probe design: 4K loci known to work for birds & reptiles, 2-3k loci in mammals, and up to 1k loci in amphibians
- Shown to be robust for resolving both shallow and deep phylogenies

Cons

- Might not work well for heavily degraded samples (historic DNA) since it requires genomic libraries with relatively large inserts (>500bp)

slide courtesy of Ke Bi

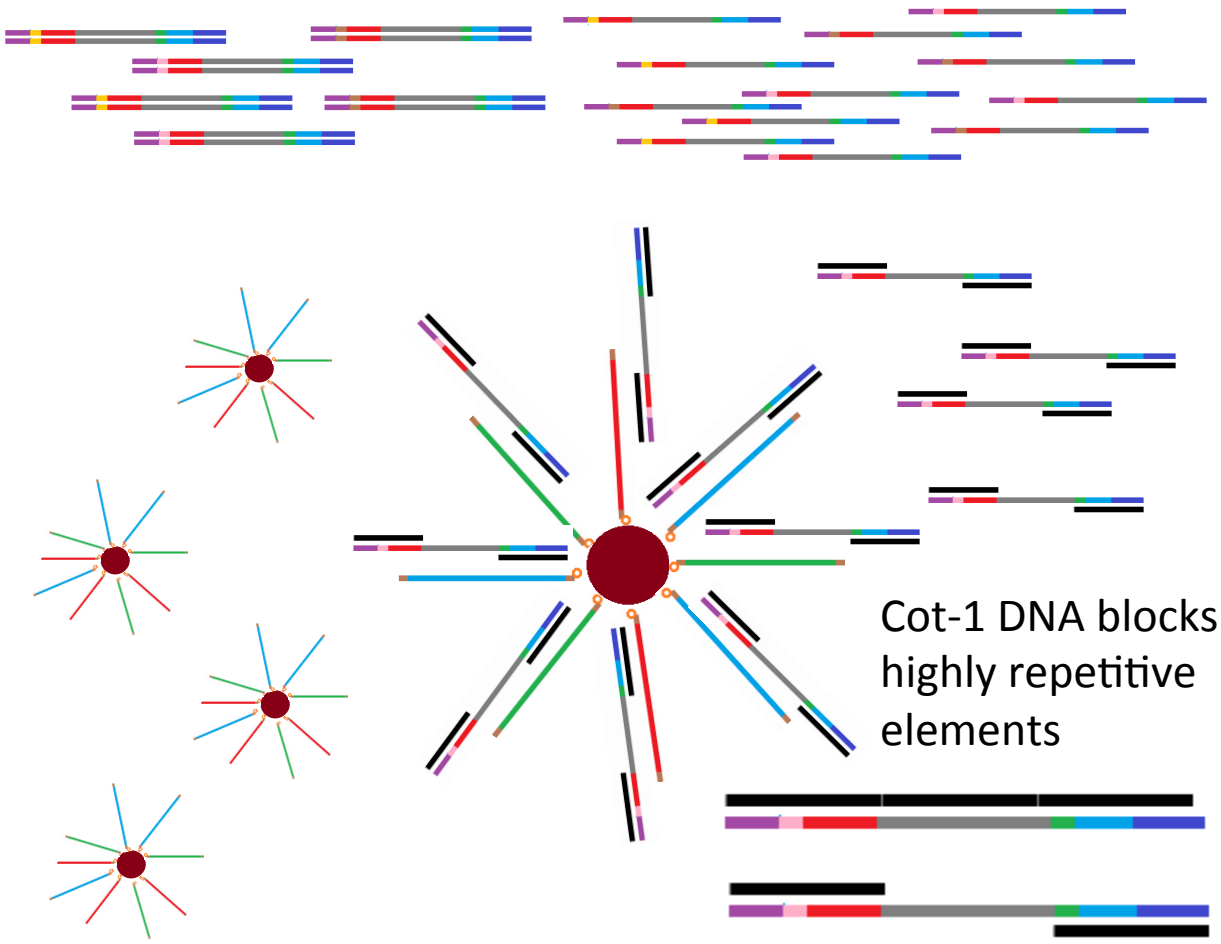
In-Solution Hybridization

Denature pooled libraries

Hybridize
With biotinylated probes and blockers
16-72 hrs at 47-65º C

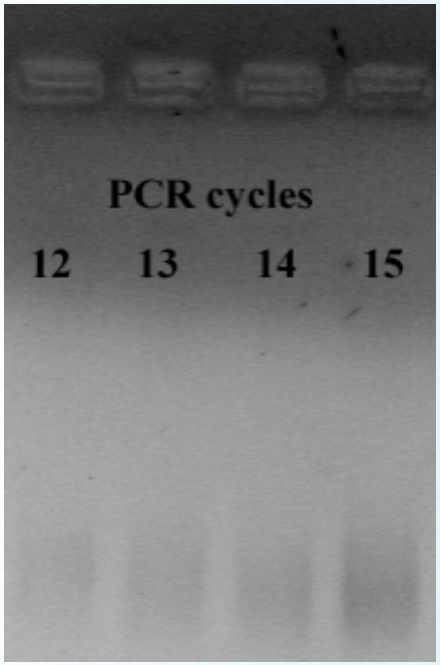
Capture
Introduce streptavidin-coated beads. Pull probes & captured libraries from solution

***Blocking oligo selection is the #1 determinant of capture efficiency**



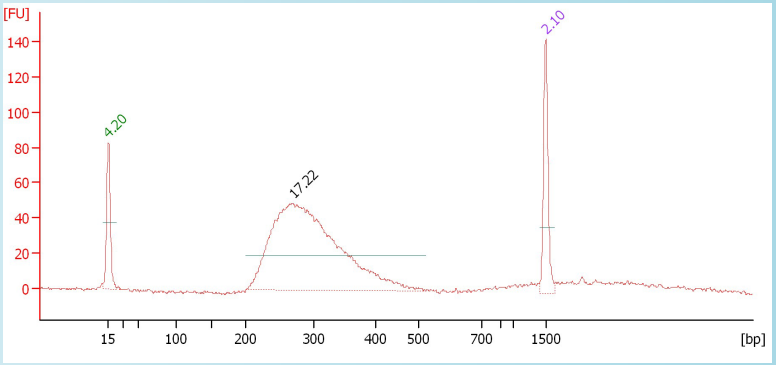
Post Capture Enrichment PCR – Avoid Over Amplification!

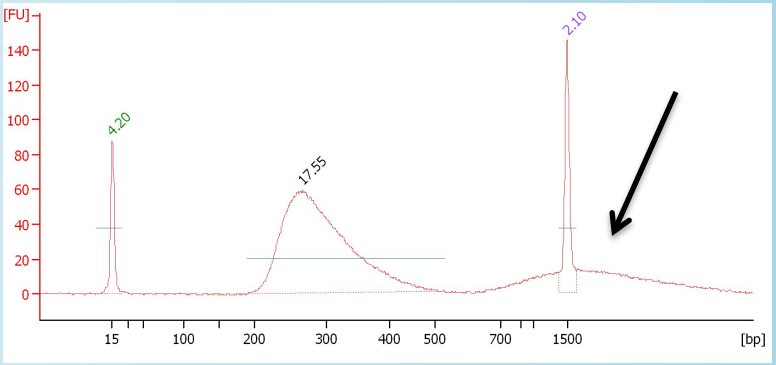
- In post capture enrichment PCR, there is a high probability of barcode swapping especially after PCR reaches saturation – short adapters can act as primers that may anneal to adapters containing different barcodes. Solution: amplify as few cycles as possible and never let your PCR reach plateau.
- To figure out how many cycles are needed – qPCR or quick PCR tests.



Nanodrop readings:
12 cycles: 12ng/ul
13 cycles: 19ng/ul
14 cycles: 28ng/ul
15 cycles: 35ng/ul

choose 12 or 13 cycles

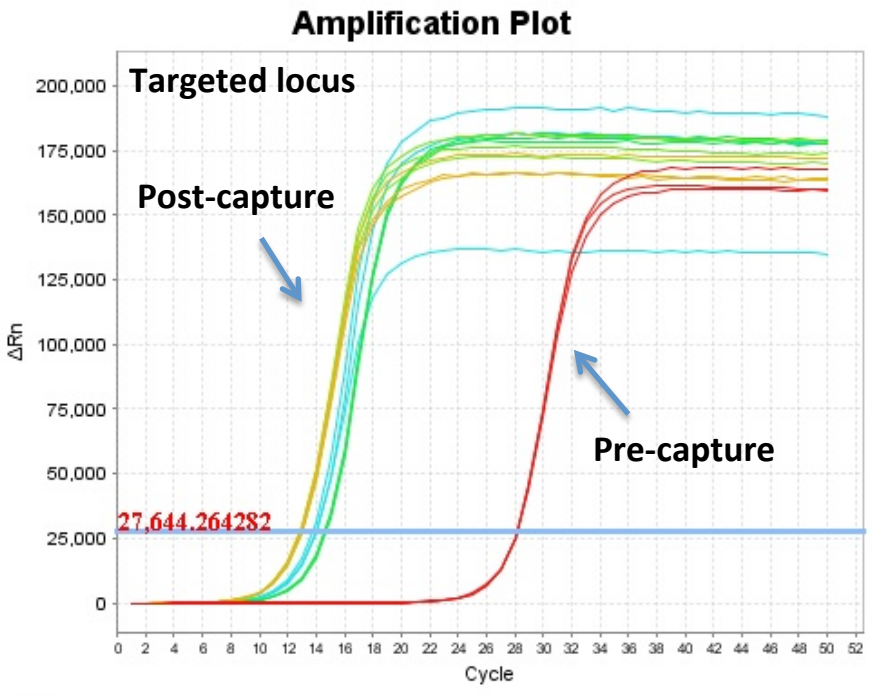




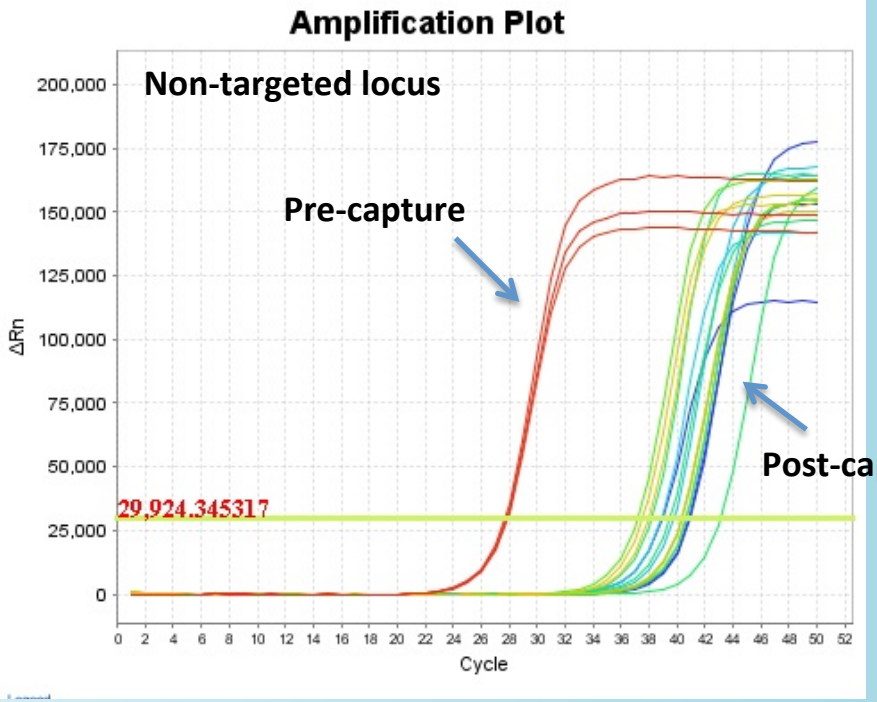
slide courtesy of Ke Bi

Estimating Enrichment Efficiency using qPCR

qPCR assays are used to estimate relative fold enrichment by measuring the relative abundance of target loci (positive assays) and non-target loci (negative assays) in pre-capture sample library and post-capture captured multiplex DNA. These assays are an inexpensive way to determine whether the capture was successful prior to sequencing.



Positive assay: enrichment of targets



Negative assay: depletion of targets

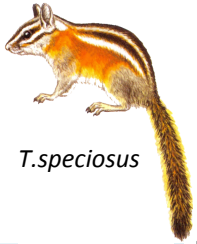
Performance of Transcriptome-based Exon Capture in a Case Study (Solution-based)

Total target size: 9.32 Mb

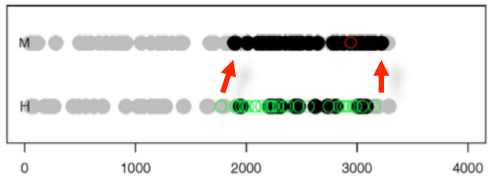
- ~2000 “candidate loci” in relevant pathways;
- 9774 assembled contigs with baits extended to their flanking regions;
- Control loci for contamination and qPCR.

Samples to survey: N = 303 + outgroups

Stable at Yosemite



T.speciosus

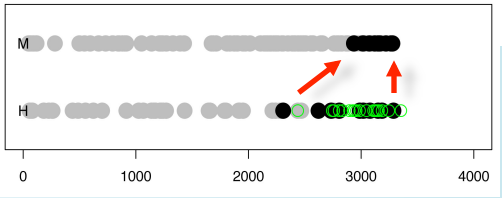


Modern: N=48

Historic: N=56

Retracting at Yosemite

T. alpinus

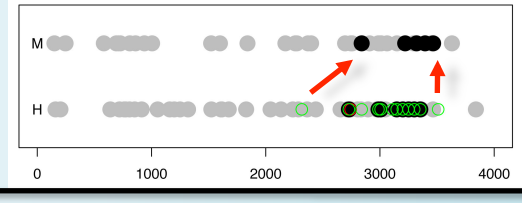


Modern: N=48

Historic: N=55

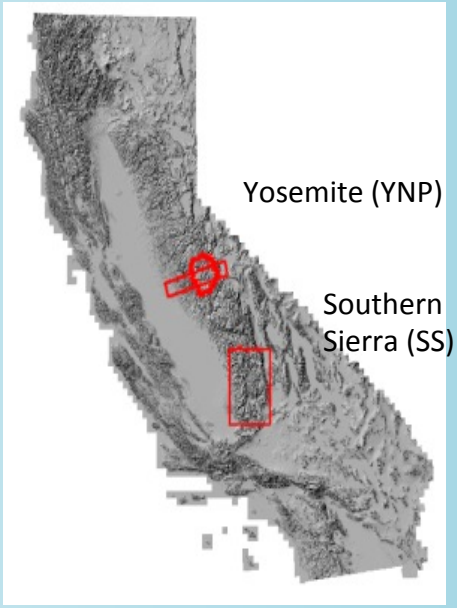
Retracting at southern Sierra

T. alpinus



Modern: N=41

Historic: N=55

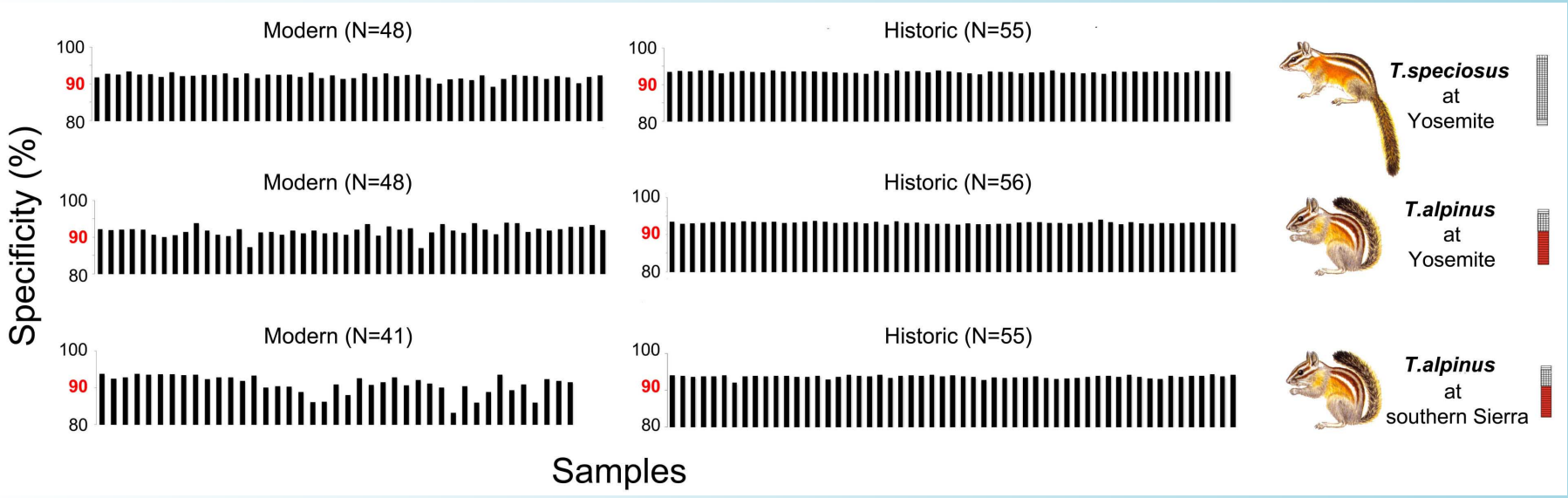


NimbleGen in-solution capture & sequencing:

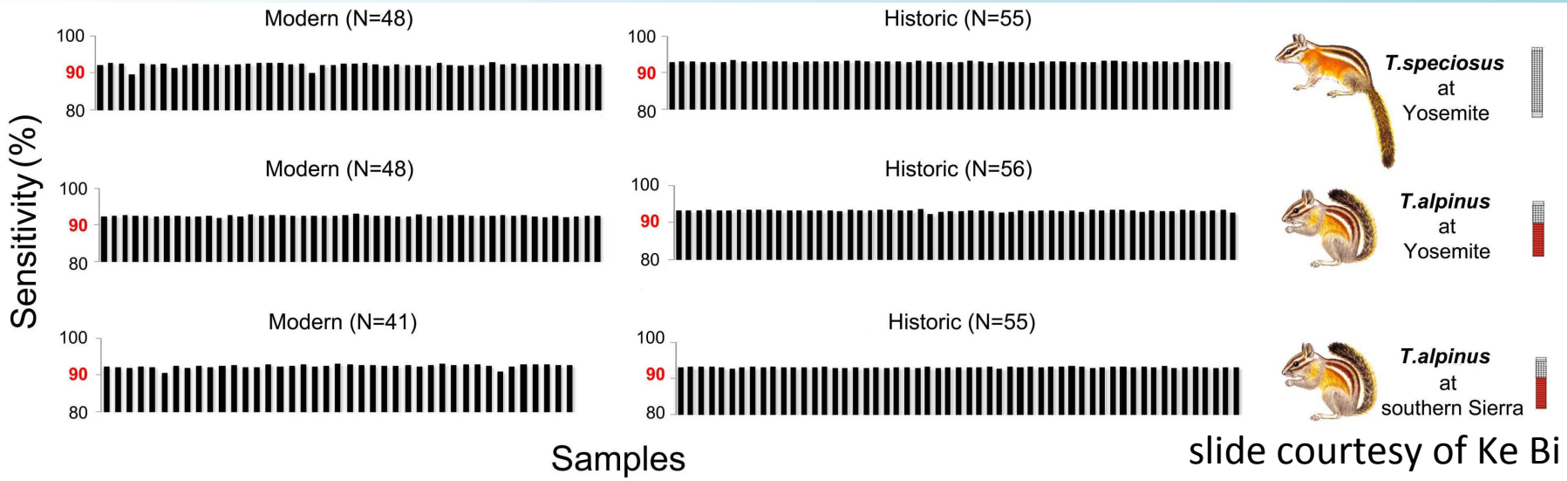
- Six capture reactions: 1 population/reaction;
- Illumina HiSeq2000, 100PE, 6 lanes: 1 population/lane.

slide courtesy of Ke Bi

Specificity - % cleaned reads mapped to the intended exons

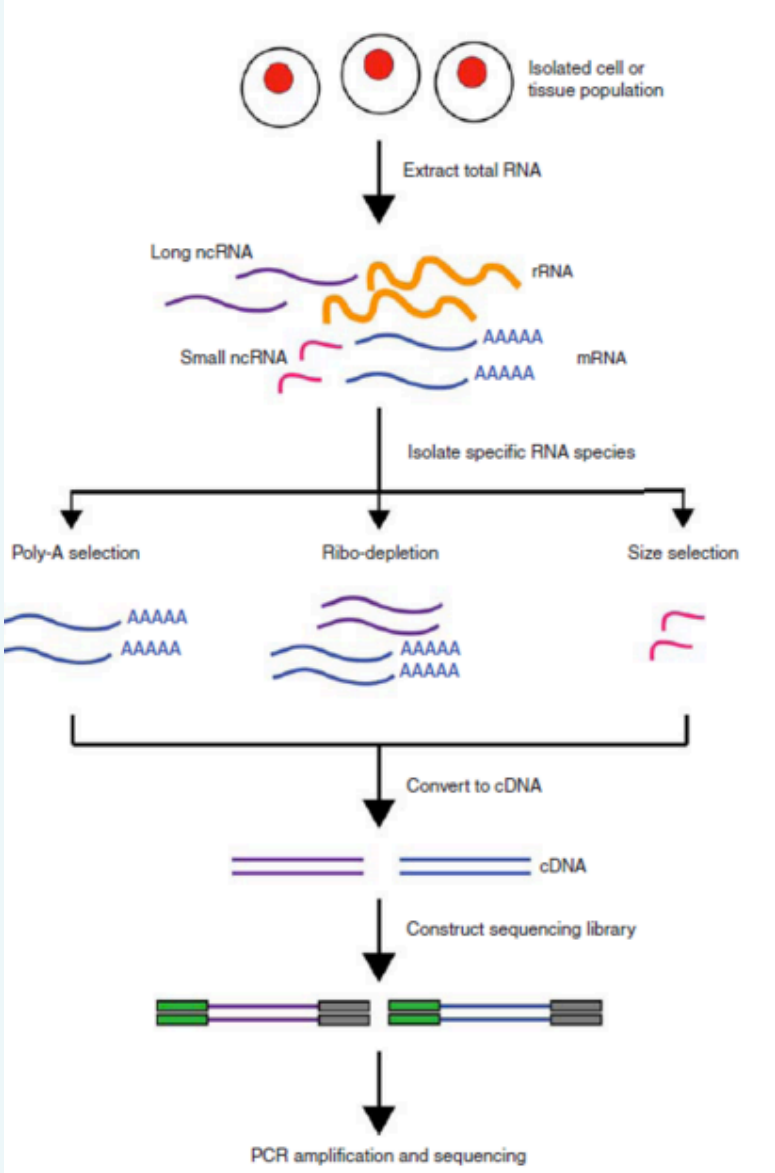


Sensitivity - % target exons represented by sequence reads



slide courtesy of Ke Bi

Illumina library preparation: RNA libraries



Kurkurba & Montgomery, 2015
doi: 10.1101/pdb.top084970

Illumina library preparation: mRNA isolation
1) Poly-A selection

- Most RNA-based projects (not all!) interested in sequencing mRNA transcripts for differential gene expression or exon sequences
- 3' poly-A tail of eukaryotic mRNA can be captured with magnetic beads attached to oligo dT probes
- Relatively inexpensive approach; standard part of most RNA library kits
- Excludes all non-mRNA RNA. However, will also exclude transcripts with degraded poly-A tails

Illumina library preparation: mRNA isolation 2) rRNA depletion

- Samples that are degraded may no longer have a viable Poly-A tail for capture
- Samples in which sequence from all non-ribosomal RNA is desired (small RNA, viral RNA, noncoding regulatory RNA)

- RNaseH digestion of rRNA hybridized to probes (human/mouse/rat)
  - NEBNext® rRNA Depletion Kit
  - Kapa Biosystems RiboErase Stranded Library Prep kit

- rRNA depletion with magnetic beads
  - RiboMinus (Thermo Fisher) (human/mouse, yeast, bacteria, plant, eukaryote)
  - RiboZero (Illumina): kit alone or library prep (human/mouse/rat**, bacteria, plant, yeast)
    - http://www.illumina.com/products/rrna-globin-mrna-removal-kit-selection-guide.html
    - **http://www.illumina.com/products/rrna-removal-kit-species-compatibility.html

- Can be up to $100 per sample! (Just for rRNA removal: does not include extraction or library prep costs) But it is well worth it when there is no other way to obtain the RNA of concern.

RNA library best practices

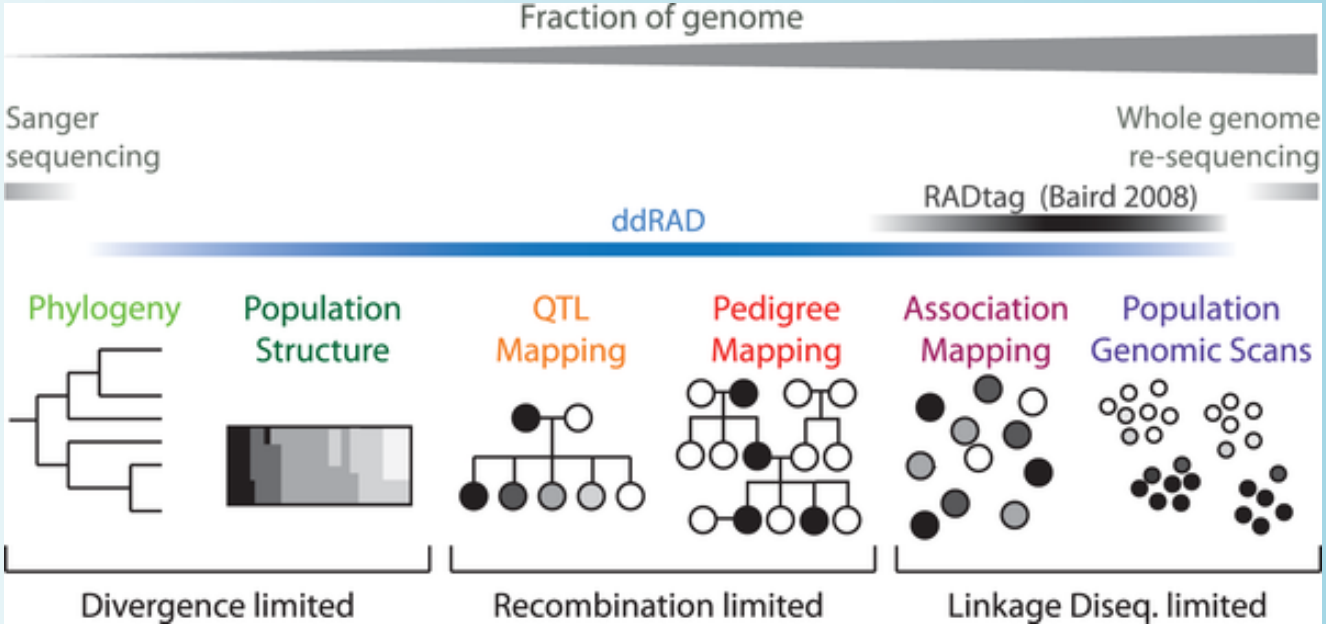
Proper RNA handling is vital prior to second strand synthesis. (After that, it is dsDNA and much heartier.)

Fragmentation step may not be exactly the same as in the protocol (8 minute standard time is often too long)

Once RNA has been converted into cDNA, the same library process can take place as for genomic DNA, post-fragmention. The only major difference is that starting material will be very low. Even greater care is required in handling and more PCR cycles may be required to have sufficient material for sequencing (often 10-15)

Illumina library preparation: Restriction site Associated DNA

- RAD-Seq can uncover hundreds or thousands of polymorphic genetic markers across the genome in a single, relatively easy & fast, cost-effective experiment without any available reference genome
- Widely used for inferring population structure, phylogeography, lower level phylogenies, introgression, trait mapping… Many case studies available



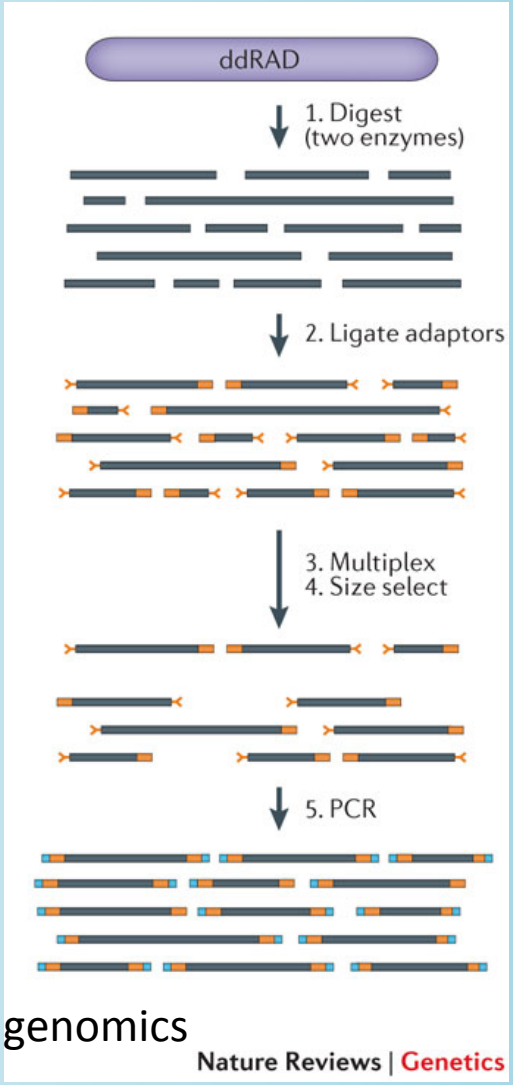
DOI: 10.1371/journal.pone.0037135

Illumina library preparation: RAD-Seq sounds amazing! what's the catch?

- Impossible to remove PCR duplicates with the most popular approach (ddRAD); must do PE sequencing of single-digest RAD
- Can take a fair amount of time to select correct restriction enzymes (RE) and to optimize. Simulation and outcomes do not always match
- Not appropriate for most phylogenetic distances due to possibility of mutations at the RE cut sites
- Big problems with allelic drop-out and high variance in coverage depth across alleles and individuals. Alleviated somewhat by ddRAD but not enough

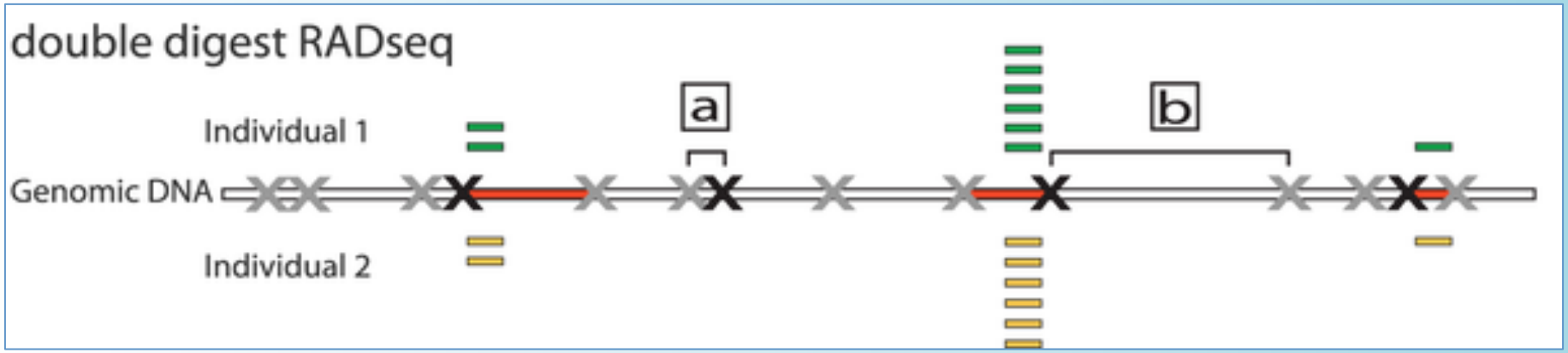
Illumina library preparation: ddRAD

- DNA is digested by two enzymes
- Adapters have overhang matching RE cut sites and internal barcodes
- Libraries are pooled and size selected with an automated instrument (Pippin Prep or similar)
- PCR amplification incorporates index and outer adapter (as in two-stage gDNA library) and enriches for fragments with the correct inner adapter combination



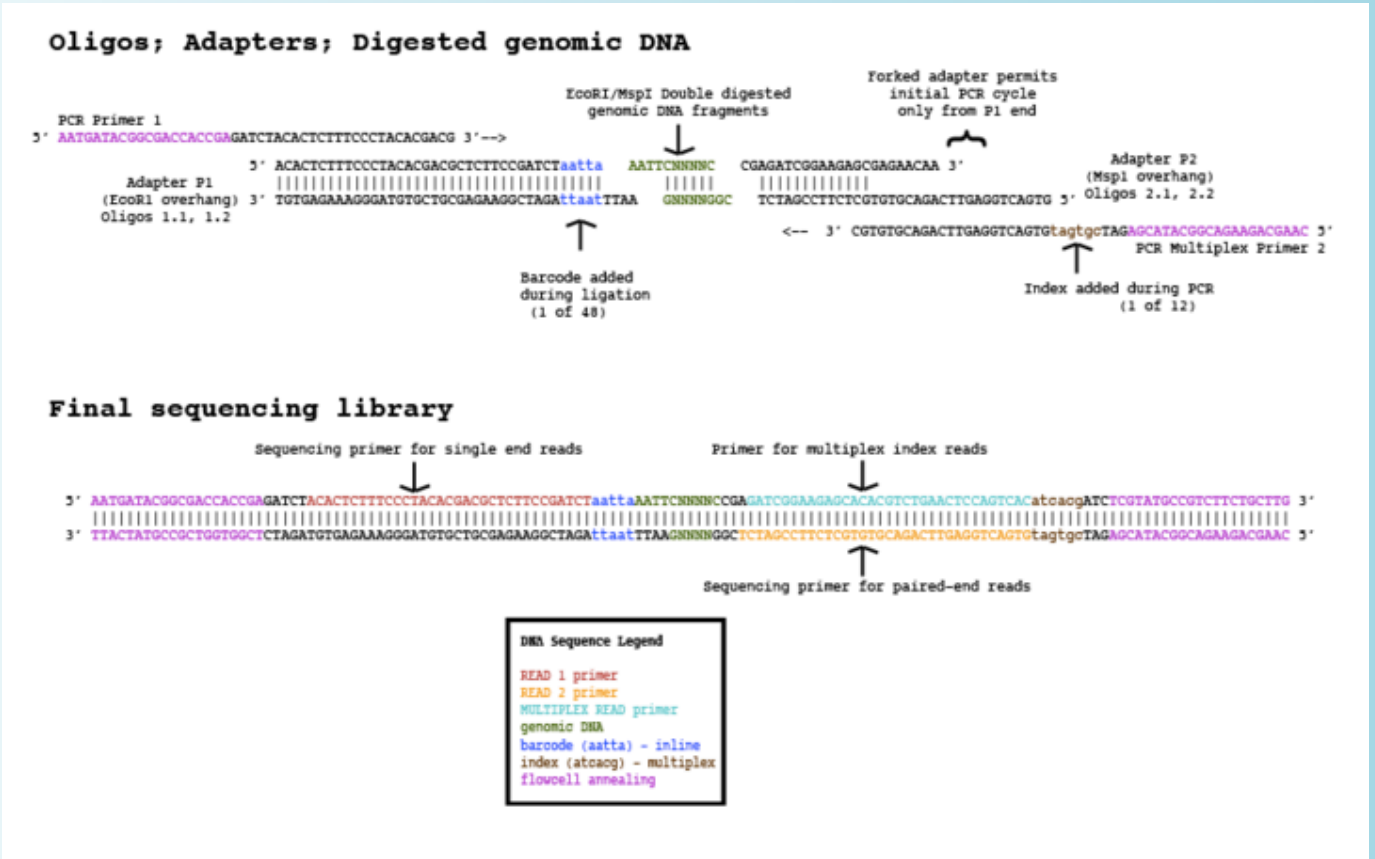
Harnessing the power of RADseq for ecological and evolutionary genomics
Nature Reviews Genetics, doi:10.1038/nrg.2015.28

Illumina library preparation: ddRAD



red regions will sequence because:

1) they are in the correct size selection range
2) they are flanked by one of each RE cut site

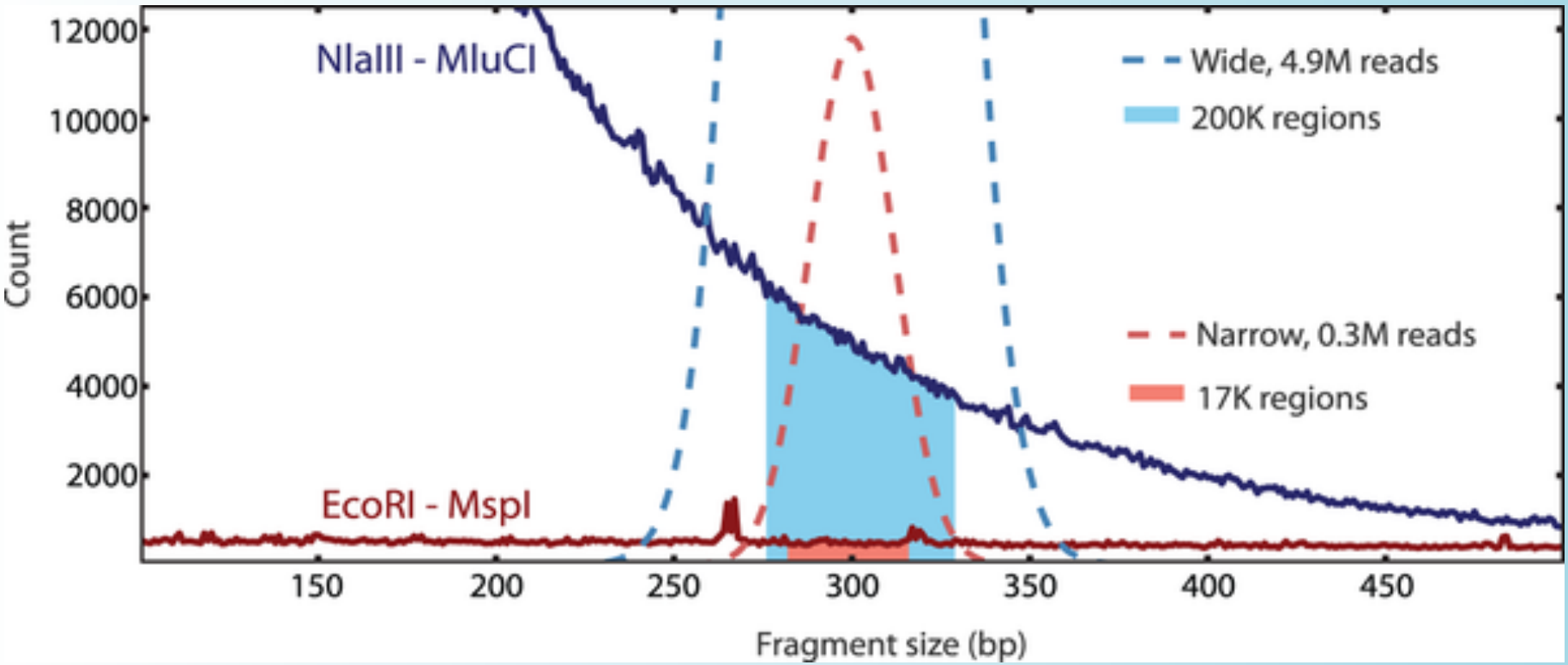


Selecting Enzymes and Size-Selection Range Considerations

- How many markers are needed/desired to answer your question?

  REs with short recognition sites (4 or 6 bps) will cut more frequently than REs with longer recognition sites (8 bps)
- Is size of genome known?
- If genome is available, can use in silico digestion like SimRAD R package (DOI: 10.1111/1755-0998.12273)
- If genome is not available, can test single and double RE digests on bioanalyzer: instructions in http://www.bit.ly/ddRAD

Figure 3. Double digest RAD sequencing provides flexibility in the number of homologous fragments recovered.



Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLoS ONE 7(5): e37135. doi:10.1371/journal.pone.0037135
http://journals.plos.org/plosone/article?id=info:doi/10.1371/journal.pone.0037135

Changing the restriction enzyme (RE) or size-selection regime modifies the fraction of genome recovered. Simulation 1 (blue lines, shading): the expected fragment size distribution for a RE digest with NlaIII and MluCI (CATG and AATT) in the Mus musculus genome is shown (solid blue line). "Broad" size selection (300 bp±50 bp) is modeled by a normal sampling distribution (mean = 300 bp, SD = 25 bp). Under this sampling distribution, 4,900,000 sequence reads (dashed blue line) are expected to cover ~119,000 regions at 7× or greater (blue area). Simulation 2 (red lines, shading): the expected fragment size distribution for a digest with EcoRI and MspI (GAATTC and CCGG) is shown (solid red line). "Narrow" size selection (300 bp±24 bp; see text) is modeled by a normal sampling distribution (mean = 300 bp, SD = 11 bp; see Analysis S1 Supporting Figure 1). Under this sampling distribution, an investment of 315,000 sequence reads (dashed red line) is sufficient to recover ~17,000 regions at 7× or greater (red area).

General RAD-Seq advice

- Resuspend/elute DNA when extracting into a no EDTA buffer (10mM Tris HCl). EDTA may impact enzyme efficiency in subsequent steps
- Be aware of available adapters; selecting restriction enzymes that match adapters already available in your lab can save a huge amount of money.
- Consult with Pippin Prep/Caliper operators to make sure the width and position of the region to cut is within the operational range of the instrument. A narrow range is not always possible and even when that can be programmed, real world results are often wider than expected. Typically 100bp is the smallest size range that can be cut.
- Prepare as many similar samples together as possible (considering adapter availability, input limitations, similar quality/quantity) for sequencing to have better uniformity. It is sometimes difficult to integrate a second set of samples into a project later on.
- ddRAD: consider the optional biotinylated adapter to remove libraries with tow P2 adapters
- As with gDNA and RNA libraries, do some tests cycles or qPCR to determine the minimum number of cycles required to have sufficient material for sequencing.
- Consult with sequencing facilities about whether a low PhiX spike-in is useful for increasing diversity at cut sites during the sequencing run

RAD-Seq advice from the experts at Cal Academy

- Genotype more samples than needed with the expectation of having to drop those with low coverage/low SNP count. If you leave these samples in final dataset, you will be more likely to reduce the number of SNPs shared across libraries.
- We have found we get significantly better sequencing results when we standardize the amount of DNA going into the adapter ligation step. We usually try for 100 ng per sample (*Note: as quantified after digestion and bead clean-up. They start digestion with 300-400ng*).
- Standardize the number of samples per size-selection pool. In cases where we have had one pool with less individuals (or less DNA), despite our efforts to put proportionally more of that sample in the final sequencing pool, our sequencing results never come out with even coverage. The bottom line is, as much as you can, try to put the same number of individuals in a pool and keep the amount of DNA going into ligation the same for all pools.
- DNA quality of samples is important, especially among pooled samples. **Do not pool samples for size-selection that have dramatically different DNA quality** (i.e HMW vs. sheared). Closely scrutinize DNA quality of samples prior to starting library prep: run on agarose gel, Nanodrop and Qubit.
- When DNA of similar quality and quantity are pooled together for size-selection, their respective concentrations can be brought to similar levels during enrichment PCR by altering the number of cycles. When pools contain both low and high quality samples, individual samples don't often PCR the same which results in some samples having much less coverage than others
- Keep number of individuals/lane conservatively low so as to ensure high coverage. Or be prepared to have to sequence an additional lane if coverage is too low.

The bottom line is, as much as you can, try to keep the amount of DNA going into ligation the same for all individuals, put the same number of individuals in a size-selection pool, and only pool individuals of the same DNA quality

General Library Preparation Advice: all methods

- Consult with someone who has used the protocol before for more specific advice
- Read the protocol carefully, many times through
- Take your time: everything relies on this step
- Plan your indexes carefully
- More is not always better. For PCR-required methods, use as few cycles as possible: enough to fully incorporate the index, to allow the library to be easily quantified, and to have material available as a back-up but no more: depends on starting material, but usually 7-9 cycles gDNA, 10-15 cycles RNA. Overamplification in library preparation can lead to an overabundance of PCR duplicates in sequence data or to index swapping if samples are pooled
- For libraries that will be used for captures, resuspend the amplified reactions in water rather than buffer since they will be concentrated before hybridization

	RNA Seq/Transcriptome	RAD-Seq	Amplicon sequencing
EGL sample prep costs	$50 for poly-A selection; $150 for rRNA depletion	very cheap after initial oligo investment for large numbers of samples (~$5-10 for reagents)	very cheap after initial oligo investment for large numbers of samples (~$5-10 for reagents)
museum/historical/ degraded samples?	no	no	no
special equipment?	none required (homogenizer/bead beater useful)	Pippin prep [available at the Functional Genomics lab in LSA]	none for metagenomic studies. Emulsion PCR or fluidics required for large, multilocus projects
reference genome/prior genomic information required?	useful but not required for initial data	useful to have a closely related reference genome for in silica RE tests, but not essential	no reference genome needed, but PCR primers for the areas of interest must be available.
major start-up costs		bar-coded adapters	adaptors and PCR primers
main benefits	RNA Seq: differential gene expression and transcript variants. Transcriptome: genomic reference and comparative exomic data	generates unbiased, genome-wide set of markers for SNP detection; degree of genome reduction is manipulatable based on RE selected.	metagentic studies (multiple organisms per amplicon)
main drawbacks	expense of library prep; only exons sequenced; RNA not always available in tissue samples; highly expressed genes dominate seq data and can make rare transcript sequences and isoforms difficult to identify	cannot be targeted to specific regions of the genome; homology may difficult to obtain between more distantly related taxa due to mutations at RE cleavage sites; problem of uniformity of results	amplicon length determined by current Illumina PE max (300 on MiSeq); very labor intensive / costly beyond a few loci

	array-based capture	in-solution capture (Nimblegen)	in-solution capture (MyBaits)
sample prep costs	$15	$15	$15
museum/historical/degraded samples?	yes	yes	yes
special equipment?	hybridization oven	none	none
reference genome/prior genomic information required?	reference genome or transcriptome required from a closely related species	reference genome or transcriptome required—may be from a more distantly related species (ex. using Xenopus exons to capture other frogs)	reference genome or transcriptome required for custom kits; none for UCE or mtDNA capture
start-up costs	array: $750 capture reagents: $150	kit: $8000 (5 reactions) capture reagents: $100/rxn	kit: $600 (8 UCE rxns) kit: $2400 (12 custom rxns) capture reagents: $100/rxn
main benefits	low per-capture cost for a lot of data (1 million custom probes); pilot projects (proof of concept); great for population genetic studies where no sample is very distant from the probe seq	large-scale capture projects; very high mapping efficiencies; 2.1 million custom probes allow very high target sizes and capture multiplexing	lowest cost in-solution capture method; very small buy-in for pre-designed kits; more capture reactions available than Nimblegen; ideal for phylogenetic studies
main drawbacks	array has technical challenges; lower mapping efficiencies than in-solution; needs reference genome or transcriptome; probes should be <5% divergent from targets; need for species-specific cot-1; arrays being phased out by vendors (still available from Agilent)	high kit prices require very large initial investment; needs reference genome or transcriptome for designing probes	Far fewer probes than Nimblegen so multiplexing is reduced (6-8 libraries max); UCEs only available for vertebrates and hymenoptera currently

Library quality control

Prior to sequencing, the following quality control checks are recommended or required:

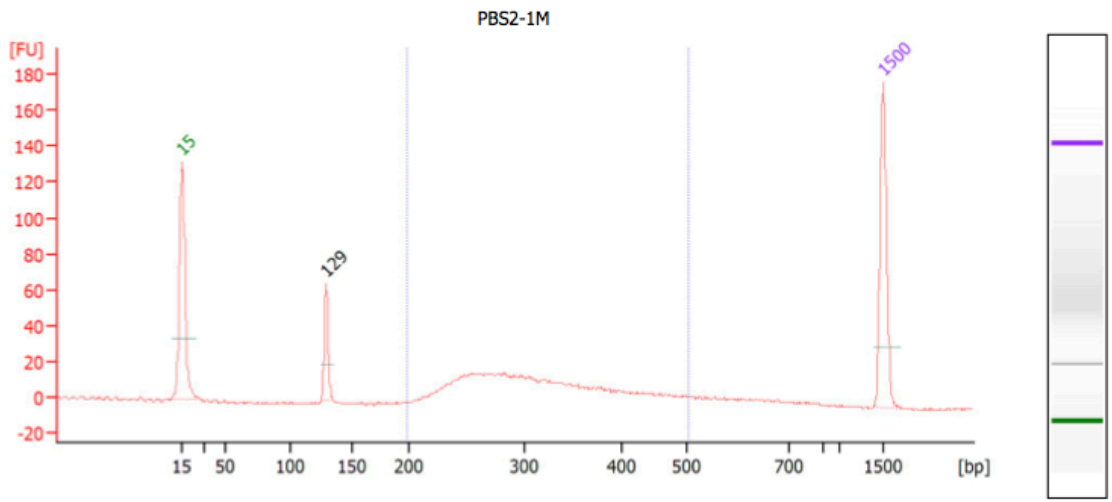
1) Bioanalyzer provides length information (required by GSL)
2) Qubit provides concentration information (required by GSL)
3) qPCR with standards provides molarity (optional)

Bioanalyzer and qubit are available at the FGL and EGL or can be run after sample submission by the GSL. I recommend assessment before submission so that any problems can be fixed in advance (samples concentrated; adapter dimer removed, "dud" samples removed/replaced.)

qPCR is usually run by GSL (one free qPCR per lane) but may be done in the EGL if a researcher wants to pool many samples his/herself.

Library quality control: bioanalyzer cDNA library

Electropherogram Summary Continued ...



Overall Results for sample 4 : PBS2-1M

Number of peaks found: 1 Area 1: 290.1

Peak table for sample 4 : PBS2-1M

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	15	4.20	424.2	Lower Marker
2	129	0.99	11.6	
3	1,500	2.10	2.1	Upper Marker

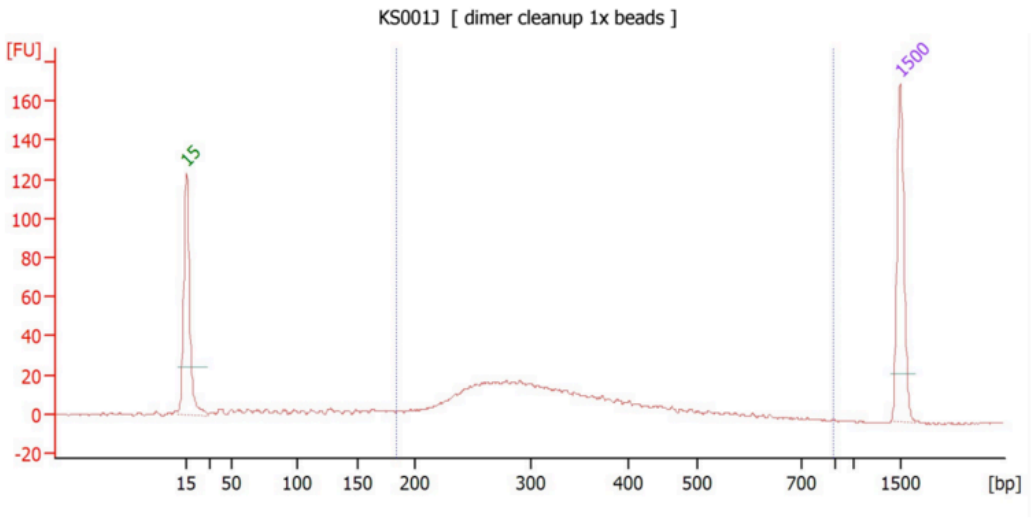
Region table for sample 4 : PBS2-1M

From [bp]	To [bp]	Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [ng/µl]	Color
199	502	290.1	78	320	21.7	6.17	

Removal of adapter dimer to run on HiSeq4000 recommended if the yield of the library can be preserved above the minimum requirement (10µL of 3nM)

Library quality control: bioanalyzer cDNA library

Electropherogram Summary Continued ...



Overall Results for sample 3 : KS001J

Number of peaks found: 0 Area 1: 365.5

Peak table for sample 3 : KS001J

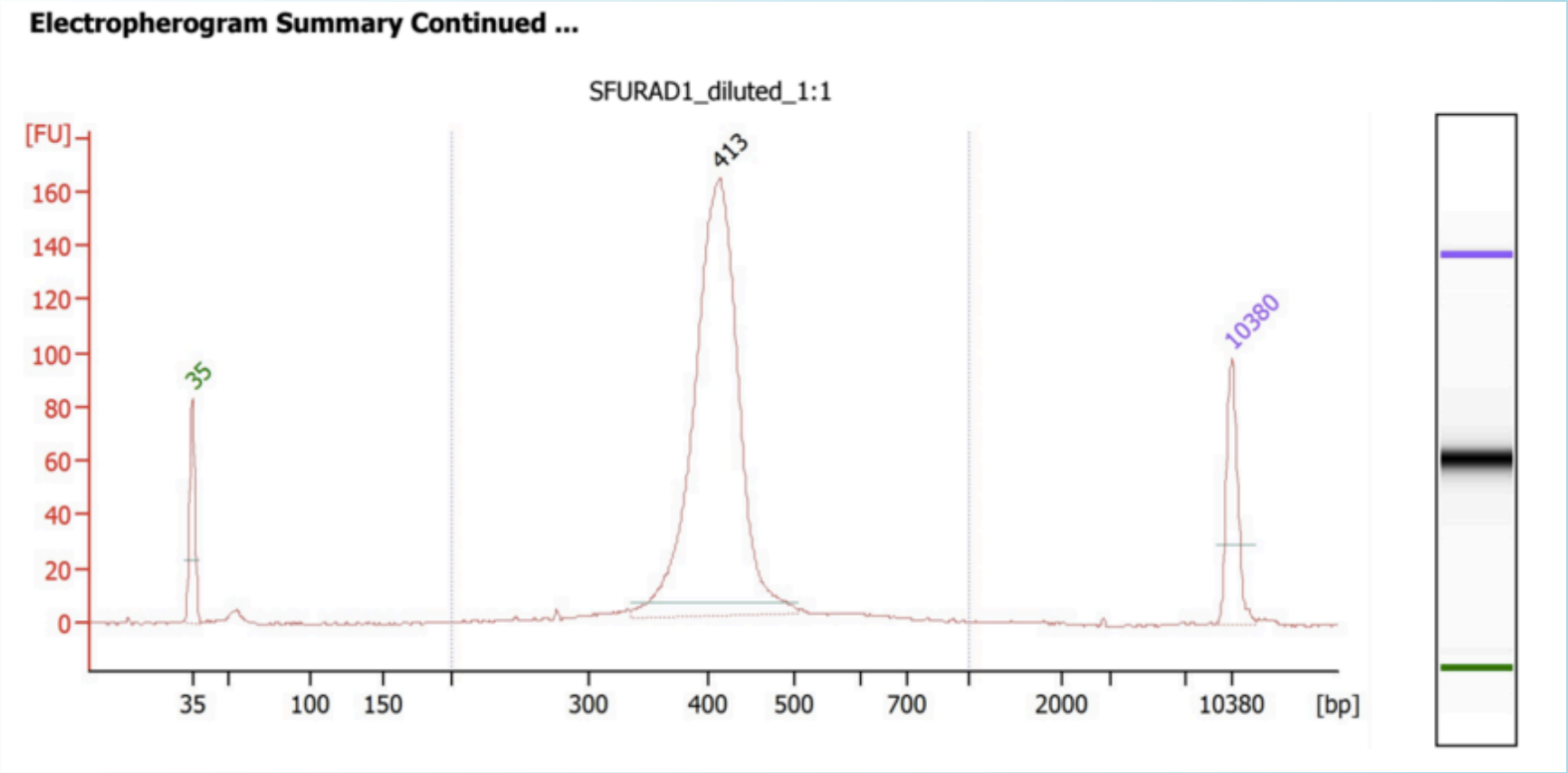
Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	15	4.20	424.2	Lower Marker
2	1,500	2.10	2.1	Upper Marker

Region table for sample 3 : KS001J

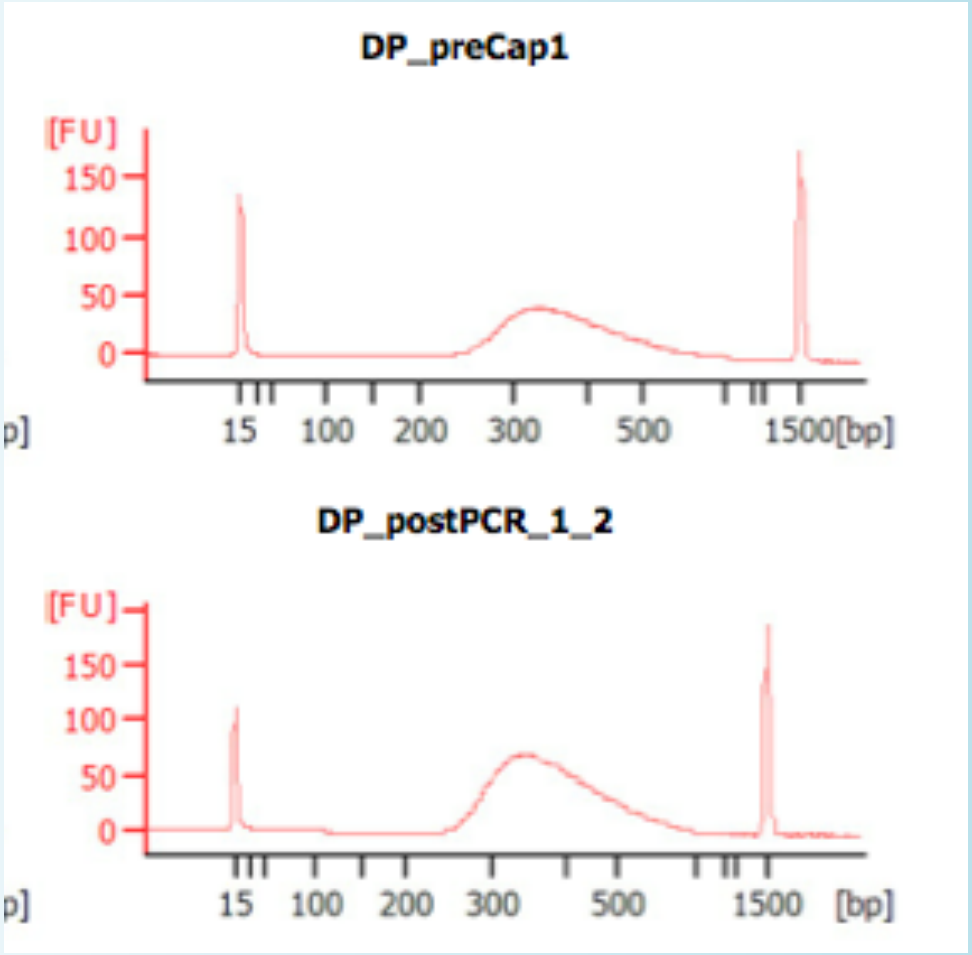
From [bp]	To [bp]	Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [ng/µl]	Color
184	844	365.5	86	335	30.1	7.86	

Same sample after 1.0x SPRI bead clean-up

Library quality control: bioanalyzer ddRAD library



Library quality control: bioanalyzer gDNA library pre- & post-capture



Library quality control: molarity estimates

Sequencing facilities usually ask for library submissions at 10µL of ≥ 10nM

The molarity of a sample can be estimated using the concentration given by the qubit and the average length given by the bioanalyzer.

dsDNA:

Formula

$$\mu g\ DNA \times \frac{pmol}{660pg} \times \frac{10^6 pg}{1\mu g} \times \frac{1}{N} = pmol\ DNA$$

N is the number of nucleotides and $\frac{660pg}{pmol}$ is the average molecular weight of a nucleotide pair.

http://www.promega.com/a/apps/biomath/index.html?calc=ugpmols

Library quality control: molarity estimates

Shortcut between concentration and molarity once average library length is known:

$$nM = \frac{\text{concentration (ng/µL)} * 10^6}{660 * \text{avg length}}$$

So, a dsDNA library of 10 ng/µL with an average length of 425bp (~300bp insert and 125bp adapter) is about 35 nM

Library quality control: qPCR with known standards

- The **best** way to know the molarity of your library is to do qPCR against known standards
- Still need bioanalyzer data for average length
- qPCR will amplify only sequence-able fragments and will include any single-stranded library fragments
- Part of standard GSL library quality control: one qPCR assessment free with every lane submission. If multiple libraries are to be pooled in a lane, they can assess multiple samples for $10 each and then pool in the desired proportions

Library submission to GSL

- Before making aliquot tubes for the GSL, place samples on a magnet in case of residual bead carry-over
- Use siliconized or low-binding microcentrifuge tubes for sample storage
- GSL Submission guidelines (read these before starting your library preps): http://qb3.berkeley.edu/qb3/gsl/submissions-seq.cfm
- Also, be prepared to acknowledge the GSL in your publications: http://qb3.berkeley.edu/qb3/gsl/faq.cfm (check with other facilities directly about the best way to acknowledge their contributions to your research.)

How many samples can be pooled together in a single lane? DNA

Warning: just an example of the thought process. Use these numbers to start talking to other people and compare your theoretical results with real-world examples from similar projects

Start with the projected amount of data:

60,000,000,000 bases

Multiply by expected (conservative) on-target percentage (say 25% for on-array capture):

15,000,000,000 bases

Divide by the target length (5Mb):

3,000

Divide by the amount of coverage you desire (be *very* conservative: this will give only the average expected coverage but it is never even across all target) (50x):

60 samples for multiplexing

How many samples can be pooled together in a single lane? RNA

Same warning as before

Often for RNA-Seq projects researchers think in the number of reads per samples (300 million for HiSeq4000)

What questions are you trying to answer?

- Are you interested in just the sequences of the 50% most highly expressed transcripts (as low as 1M reads/sample: doi:10.1016/j.gene.2014.12.013)?
- Are you interested in comparing differential expression (~30M reads/sample)
- Do you want to discover novel elements, perform more precise quantification, especially of lowly expressed transcripts (100-200M PE reads/sample or even more)

Recommended to always pool all samples together (DOI: 10.1534/genetics.110.114983) then spread the pool out over multiple lanes in order to negate to influence of lane effects. Pre-plan to have sufficient indexes available.

http://rnaseq.agbioinfo.utk.edu/images/2/20/Transcriptome_Project_Design.pdf
https://genome.ucsc.edu/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf
https://www.msi.umn.edu/sites/default/files/RNA-Seq_lecture1_0.pdf

HiSeq 4000 sequencing

- Reliably 300,000,000 clusters passing filter each lane (for PE100 = 60 gigabases of data!)
- Patterned flow cell
- Does not perform as well as the other Illumina instruments (HiSeq 2500, MiSeq) for reads > 100bp
- Not recommended amplicons
- RAD-Seq okay with a ~15% PhiX spike-in

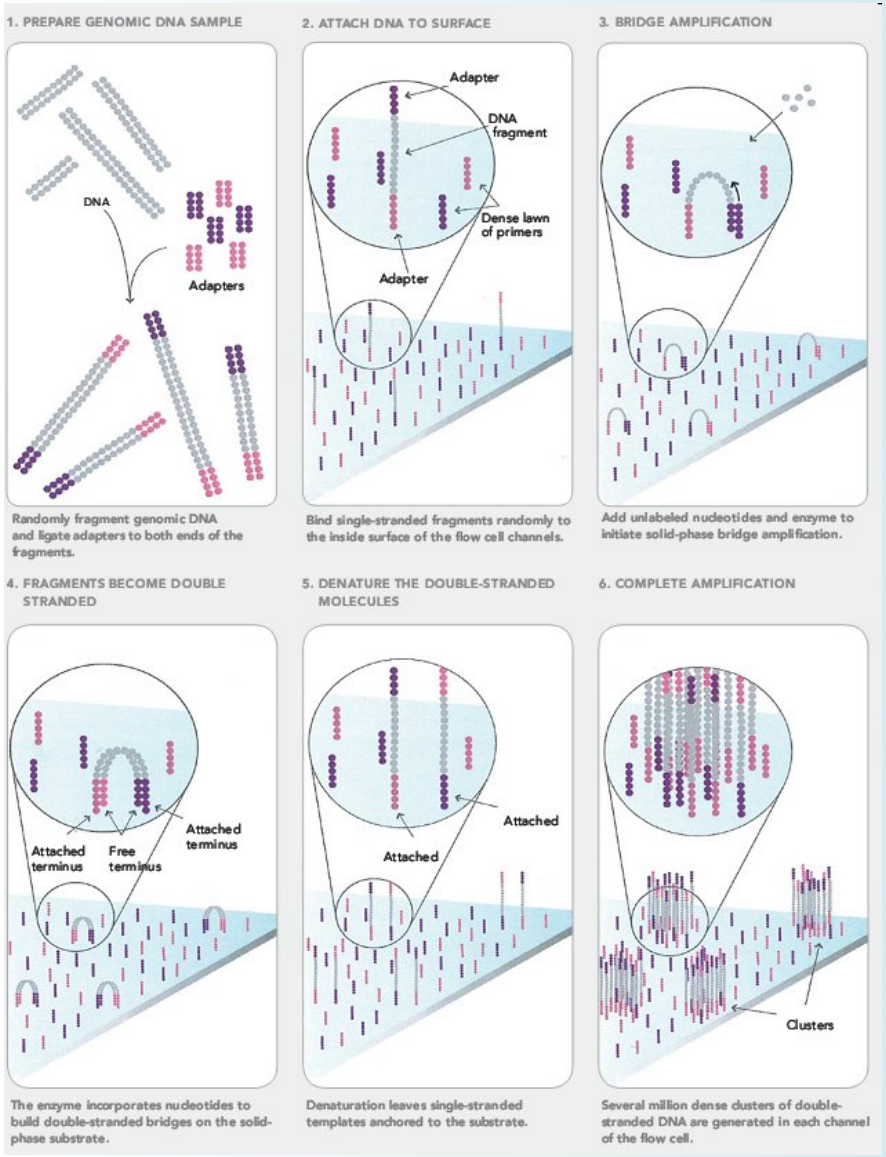
HiSeq 4000 sequencing restrictions

- Less than 0.5% total adaptor dimer
- Tightly distributed insert sizes (unless the user accepts a sequencing bias towards small molecules)
- Average library size around 470bp max. Majority of library size under 670bp (except for small tails) [not a good option for UCEs]
- Submission > 3nM, as measured via qPCR.
- Illumina compatible adaptor sequences (no custom primers)

HiSeq 2500 & MiSeq sequencing

- ~120 million clusters passing filter (HiSeq 2500) or ~25 million (MiSeq)
- More flexible in terms of library type, insert size
- Can sequence low complexity amplicon libraries
- Can sequence up to 250bp (HiSeq 2500) or 300bp (MiSeq), paired-end. (Note 300bp PE is available on the 2500 as as special request. Ask for pricing.)
- Sometimes have faster run speeds and queues than the 4000 (only 2 lanes per flowcell for the 2500 and 1 for MiSeq)

Illumina sequencing: how does it work 1



Lib Prep

Wash over Flowcell

Adapter Bridging

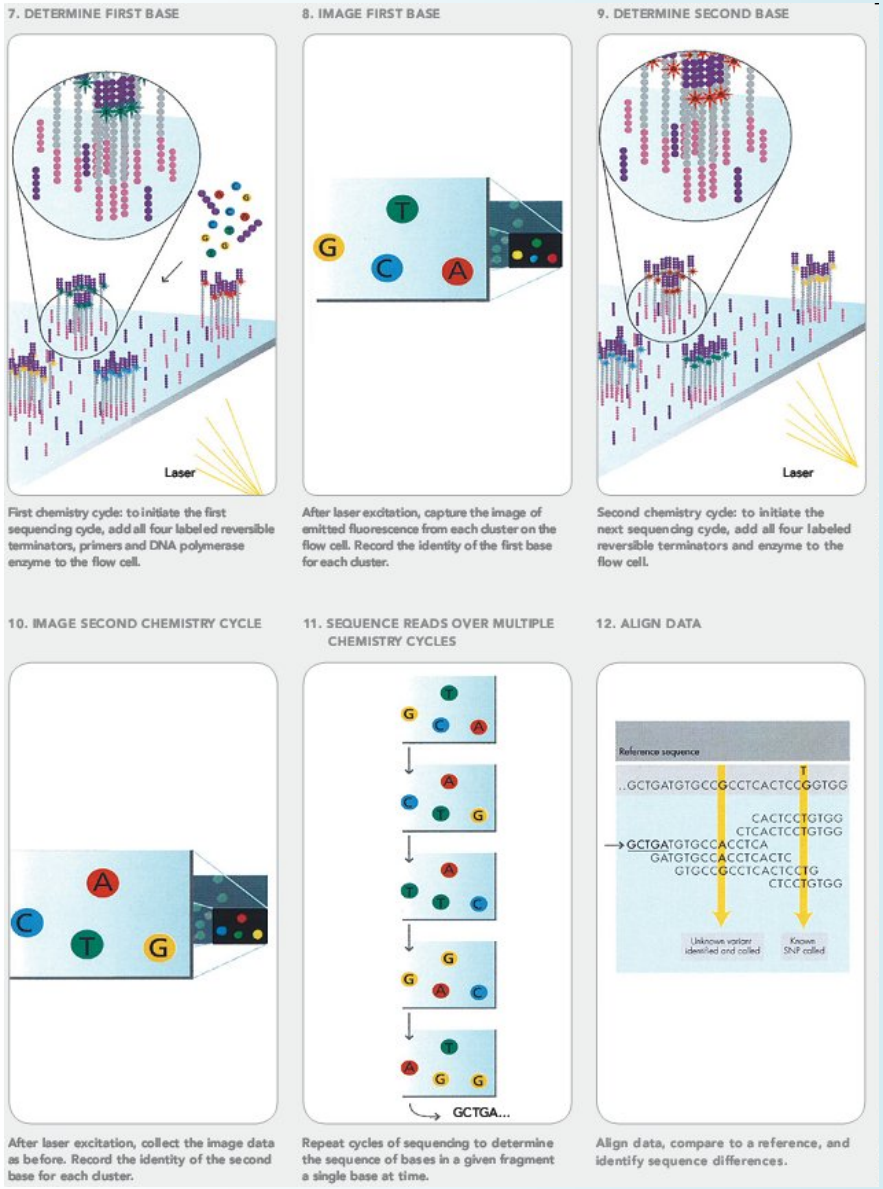
Bridge Amplify

Unbind

Cluster Formation

Detach and wash 1 adapter

Illumina sequencing: how does it work 2



Bind Seq Primers + Labeled dNTP reversible terminators

Flourescence Washed (first base read)

Labeled dNTP reversible terminators

Flourescence Washed (second base read)

Cycle con't 50bp/ 100bp/ 150bp

DATA

PAIR ENDED: wash away clusters, start from beginning, wash away different adapter before sequencing

PacBio Sequencing

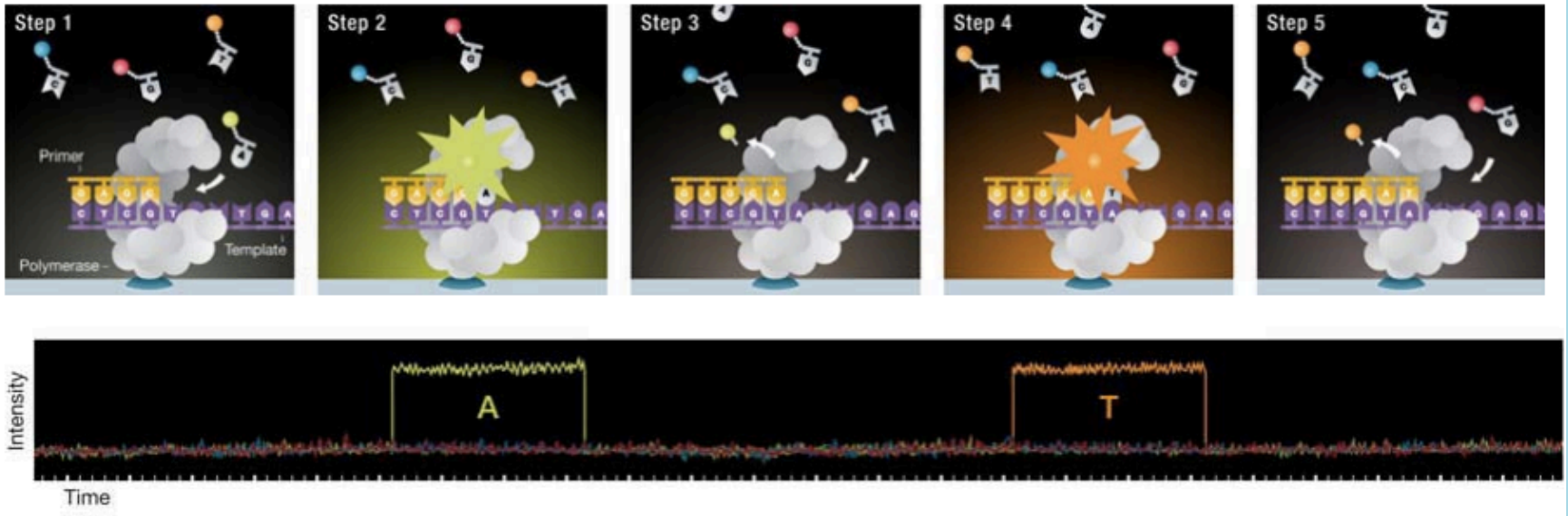
Long read sequencer available at UC Davis. GSL will help make libraries and will hopefully have funding for an instrument here in the near future.

Whole genome sequencing, whole transcript

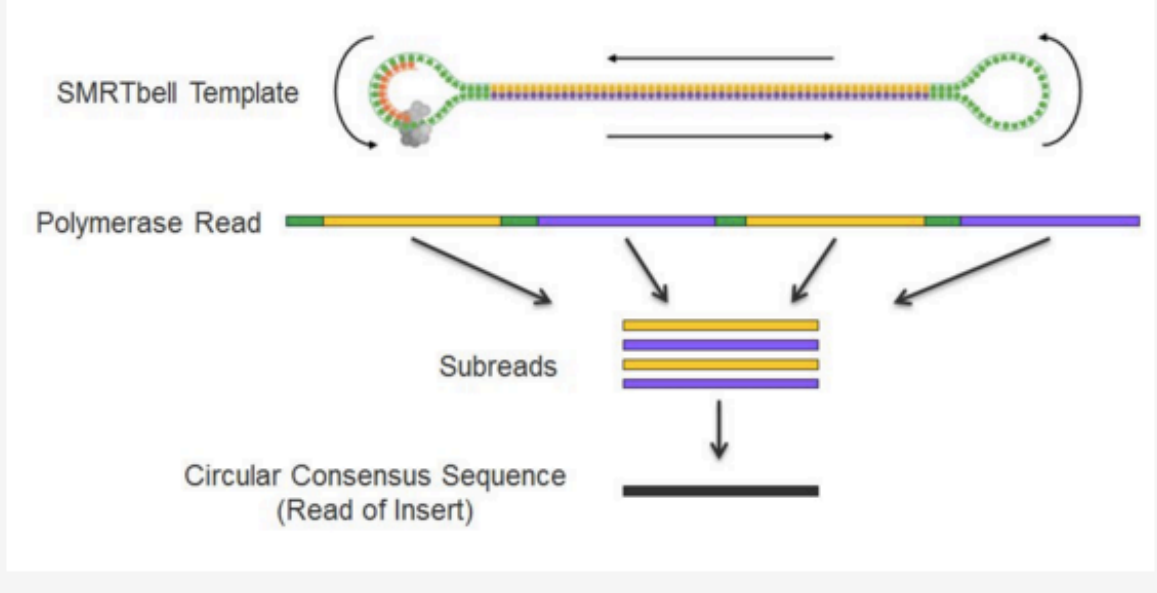
For most projects, only a few libraries are needed. Preps are expensive (> $400); best to send to a core facility with experience (Berkeley, Davis)

Not the only option for longer or linked reads (Moleculo, 10xGenomics)

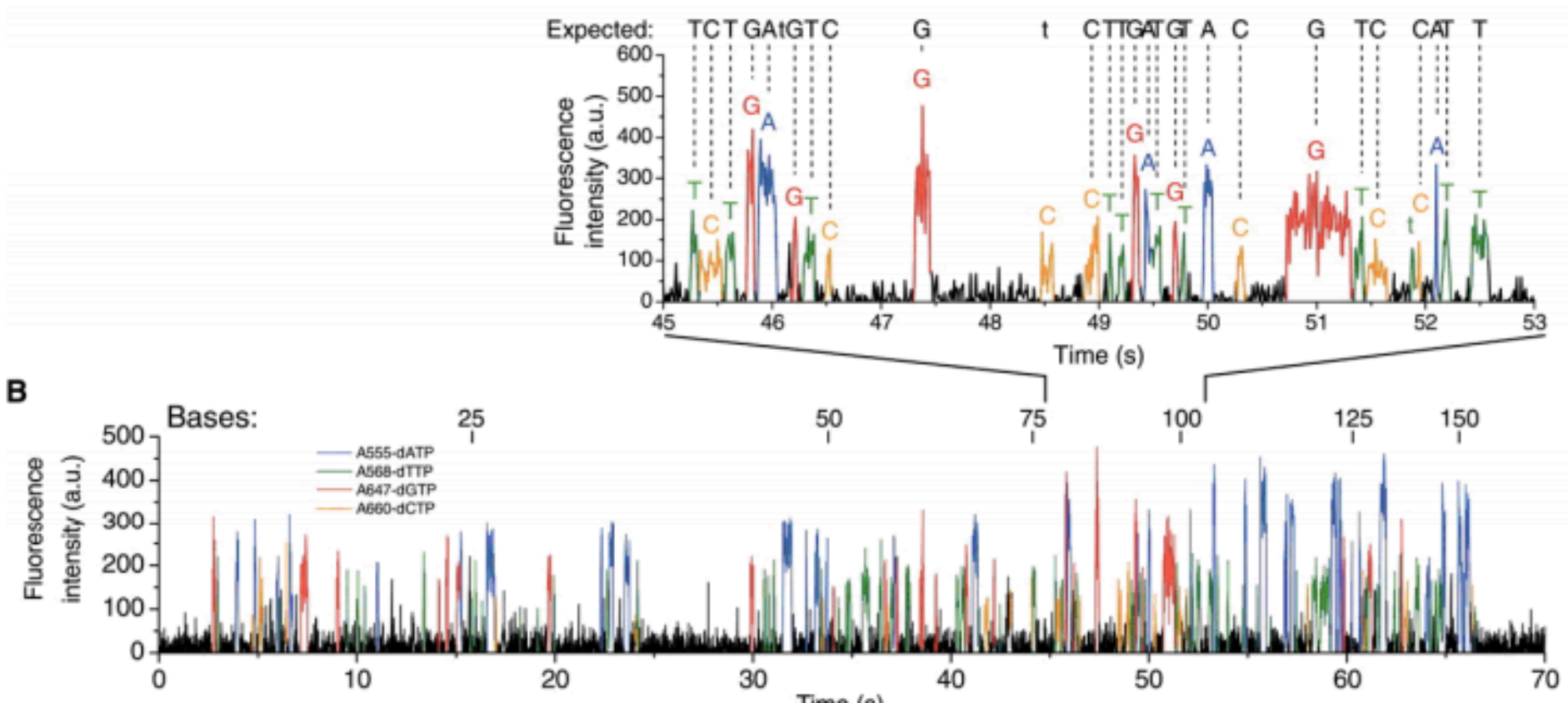
PacBio Sequencing:
Single Molecule, Real-Time (SMRT) Sequencing



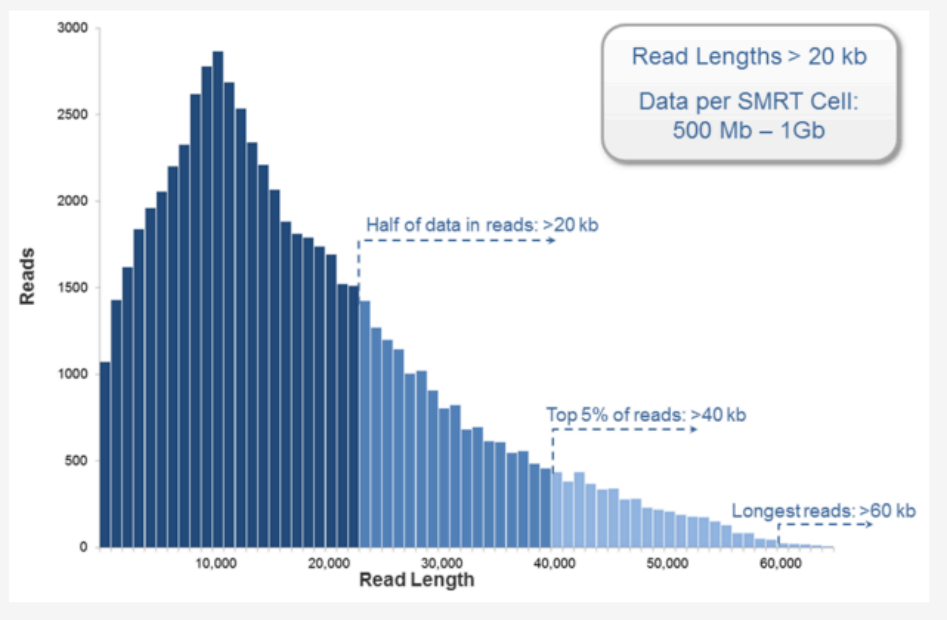
Step 1: Fluorescent phospholinked labeled nucleotides are introduced into the ZMW.
Step 2: The base being incorporated is held in the detection volume for tens of milliseconds, producing a bright flash of light.
Step 3: The phosphate chain is cleaved, releasing the attached dye molecule.
Step 4-5: The process repeats.



Pac Bio Technology

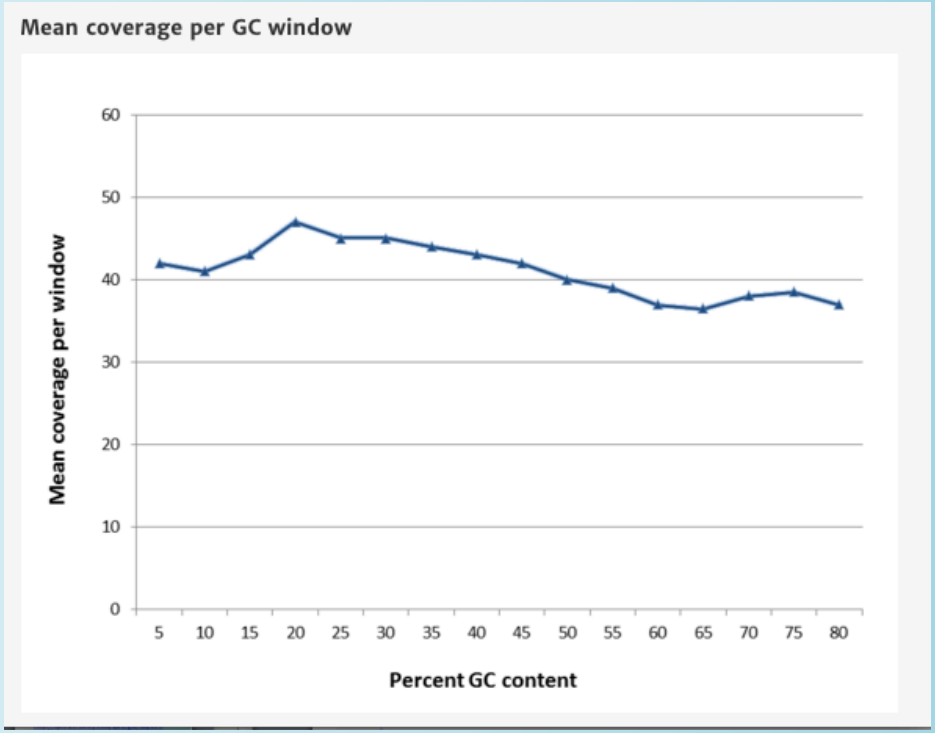


PacBio Sequencing: Advantages



1) Very long read lengths!
   Easily 10kb, up to 50kp

2) Single molecule sequencing:
No amplification biases



3) Has fewer difficulties with repetitive elements, high-GC content, homopolymers, and other challenging genomic regions

PacBio Sequencing: Disadvantages

Very high error rate: long reads that pass through the same insert multiple times reduce that to merely a high error rate (indel errors). Can be corrected with less expensive short-read data

More expensive per megabase than Illumina

Library prep is more expensive than one run: multiplexing is now supported, but the costs only make sense for most projects if large amounts of data are collected per library

Superb fairywren genome (1.2Gb: 20x coverage)

Required 80µg of DNA per library prep (20µg per SMRT cell). **Best quality DNA is essential**

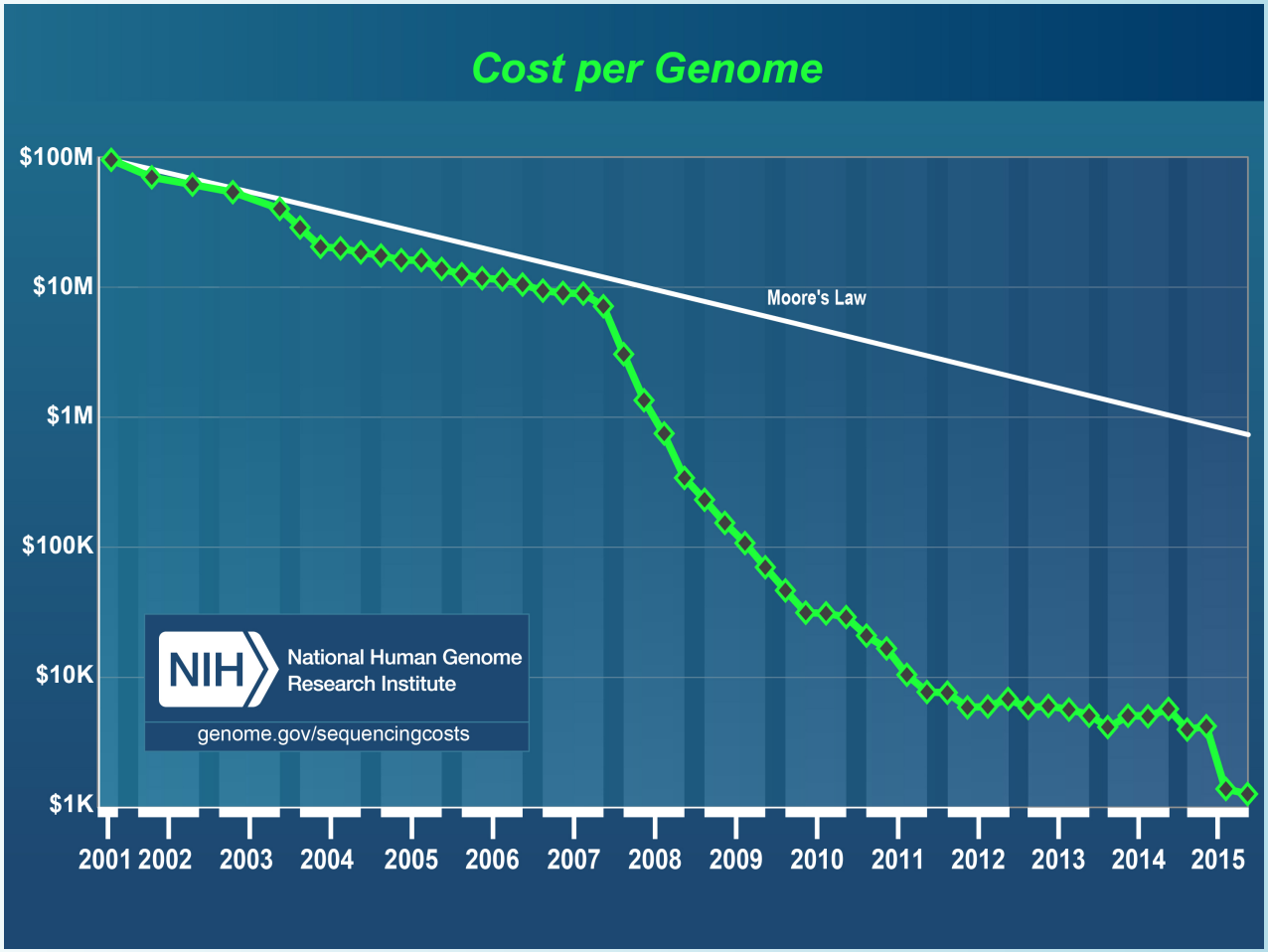
27 SMRT cell runs: mean read length = 9000bp (very consistent: 7K-10K), mean number of reads = 99,000 (varied from 18K-122K!), average of 870 megabases/run

Fewer ultralong reads than PacBio advertises (very few > 30kb) but more reads overall

About $10,000 total cost for this aspect of the project

Also used 1 HiSeq lane of 100PE and 1 MiSeq lane of 300PE for genomic data and error correction (~$5000)

Decline in sequencing costs: pattern



https://www.genome.gov/sequencingcosts/

Shana McDevitt's key advice

GSL director: shana.mcdevitt@berkeley.edu

Mailing list: gslusers@lists.berkeley.edu

- indexes matter
- understand base balance (low-complexity libraries)
- make sure you know what you will do with your analysis BEFORE you begin
- contact her or GSL staff BEFORE submitting your first libraries
- carefully follow GSL submission guidelines to receive the best quality data

In Summary

- Project design & molecular work is the foundation for everything that follows: the best bioinformatics tools cannot salvage poor data
- Consult early; consult often
- Plan out as much of your project as possible before even starting extractions; when in doubt, choose the option that provides the most flexibility to protect yourself from unforeseen changes
- Budget is important but should not be the primary factor in study design; collecting the wrong data for your study questions will not lead to a successful project no matter how cheap
- High quality nucleic acid extractions are the key to high quality libraries which are the key to high quality data
- Take your time with the lab work; better to do it slowly and correctly than rush and waste time/money

CGRL Resources

http://cgrlucb.wikispaces.com/
(also has archives of slides & resources from past presentations

4. RNAseq workshop using Galaxy and edgeR 2/18/2016
5. Genome assembly workshop part I:
Project Design, Sequencing Technologies and Library Methods. 2/22/2016
6. Genome assembly workshop part II: Bioinformatics. 2/23/2016
7. Introduction to R 3/7/2016
8. ChIP-seq data analysis 3/14/2016
9. RNAseq workshop non-model organism 4/4/2016
10. 16S amplicon sequencing workshop 4/18/2016

dates are subject to change: sign up for cgrl-announce@lists.berkeley.edu to receive workshop announcements