Applications of Sequence Captures in Non-model Organisms

Ke Bi

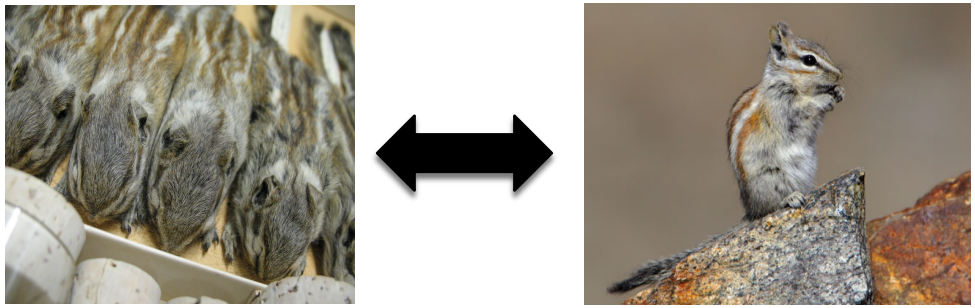
Computational Genomics Resource Laboratory (CGRL)

California Institute for Quantitative Biosciences

UC Berkeley

November 5, 2014

Genome Comparison of Museum Collections to Detect Micro-Evolutionary and Demographic Response to Recent Climate Change



Questions:

Detect genetic signature of positive selection => coding sequences

Demographic inference => non-coding sequences

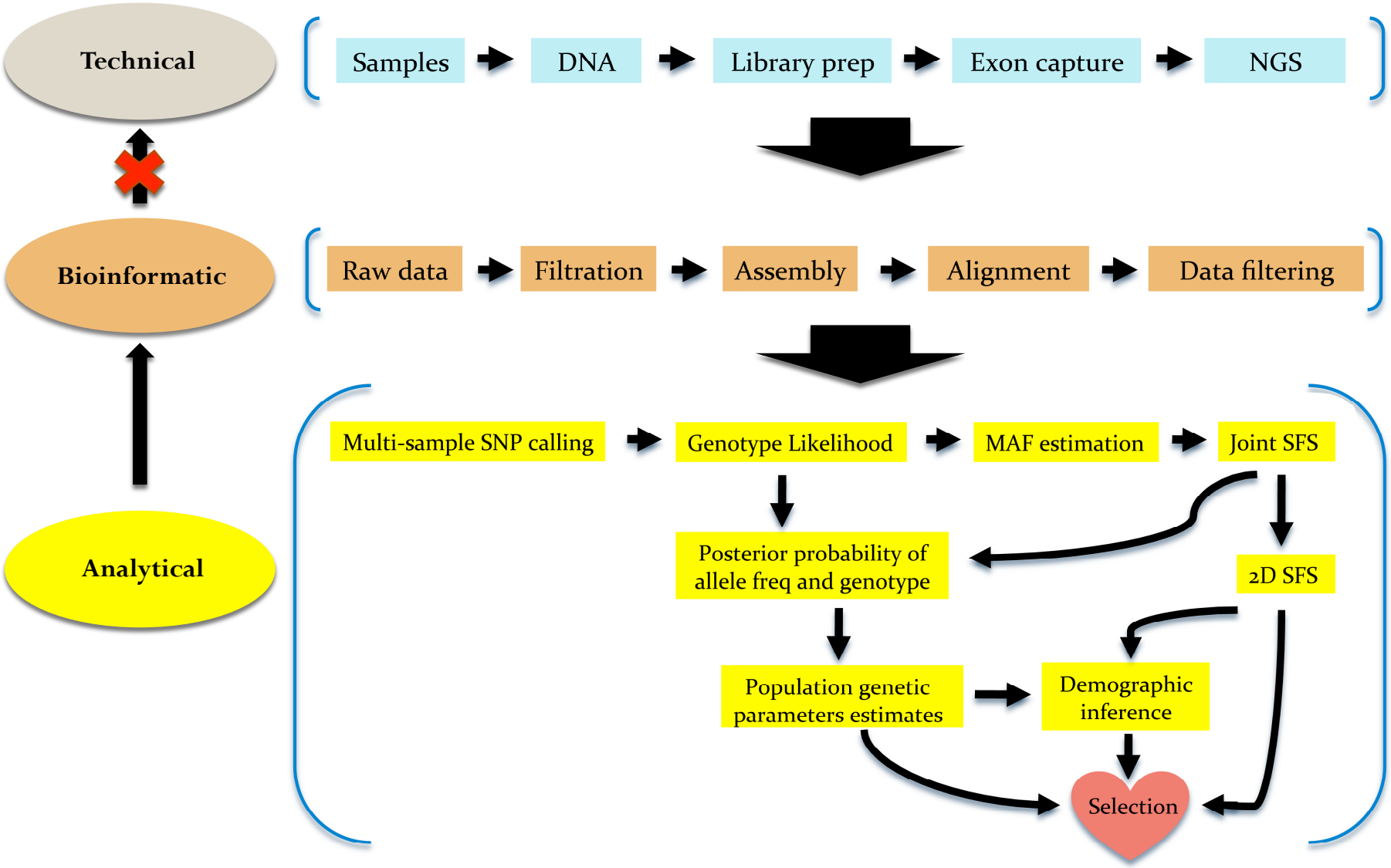
Samples:

Historic vs Modern (time series data)

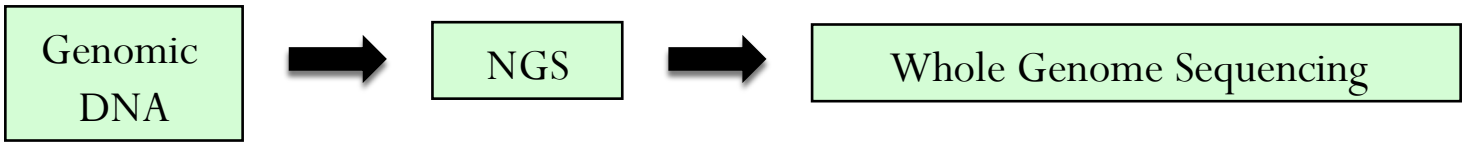
Approaches:

Exon capture + NGS for a population-level, genome-wide scan

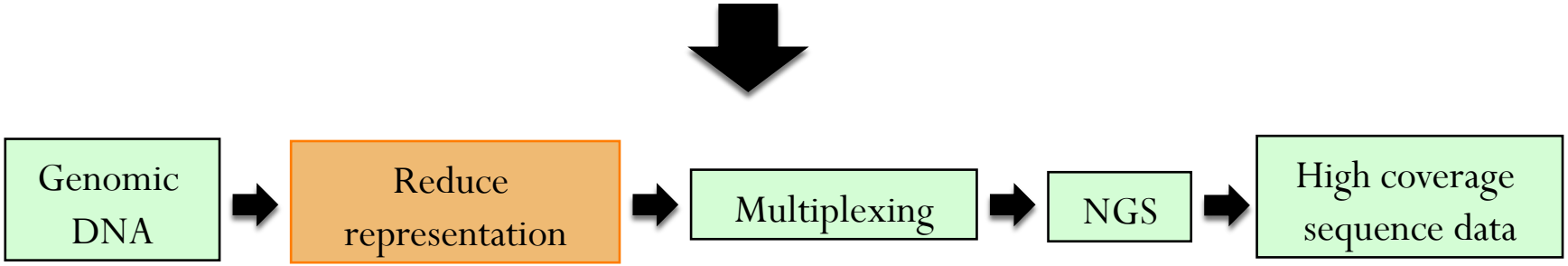
General Workflows



Marker Development in Non-model Organisms for Population Genomic and Phylogenomic Applications



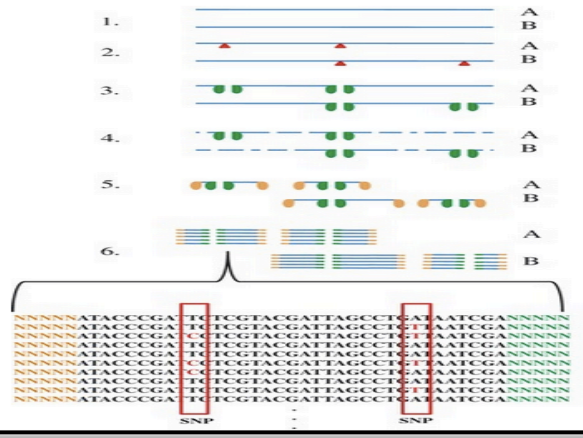
- Require a reference for alignment
- Challenging for organisms with large genomes and/or genomes that are highly repetitive (e.g. amphibians)
- Not cost-effective when sample size is large
- Not necessary for most phylogenomic and population genomic questions
- Not feasible when your budget is tight



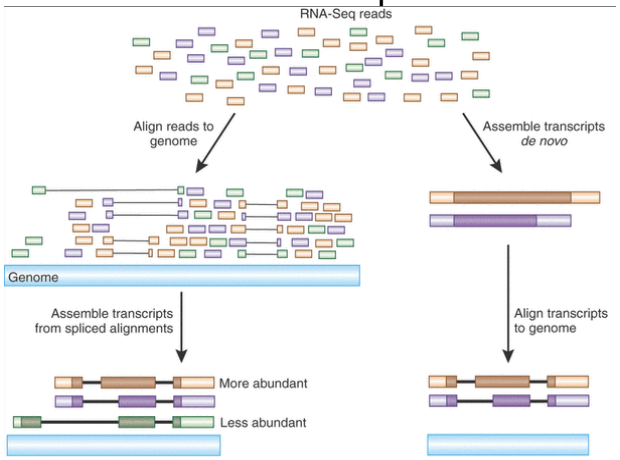
Methods for Reducing Representation in Population Genomic and Phylogenomic Studies (Genotyping by Sequencing)

GENERAL

RAD-tag Sequencing

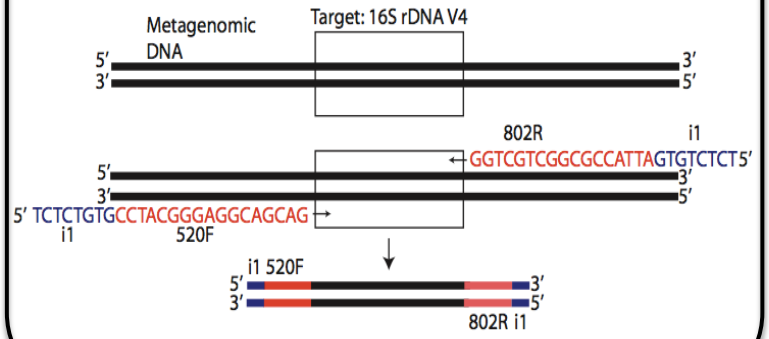


RNAseq

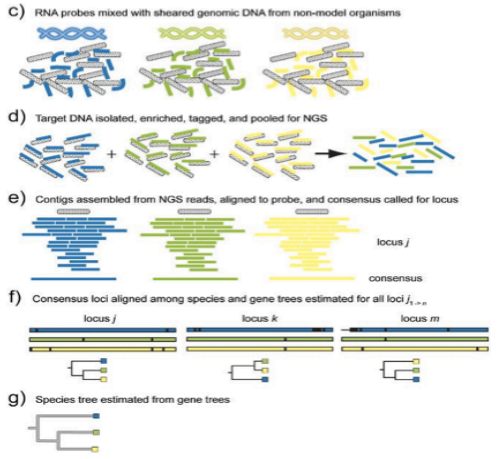


TARGET ENRICHMENT

Multiplexing PCR



Sequence captures

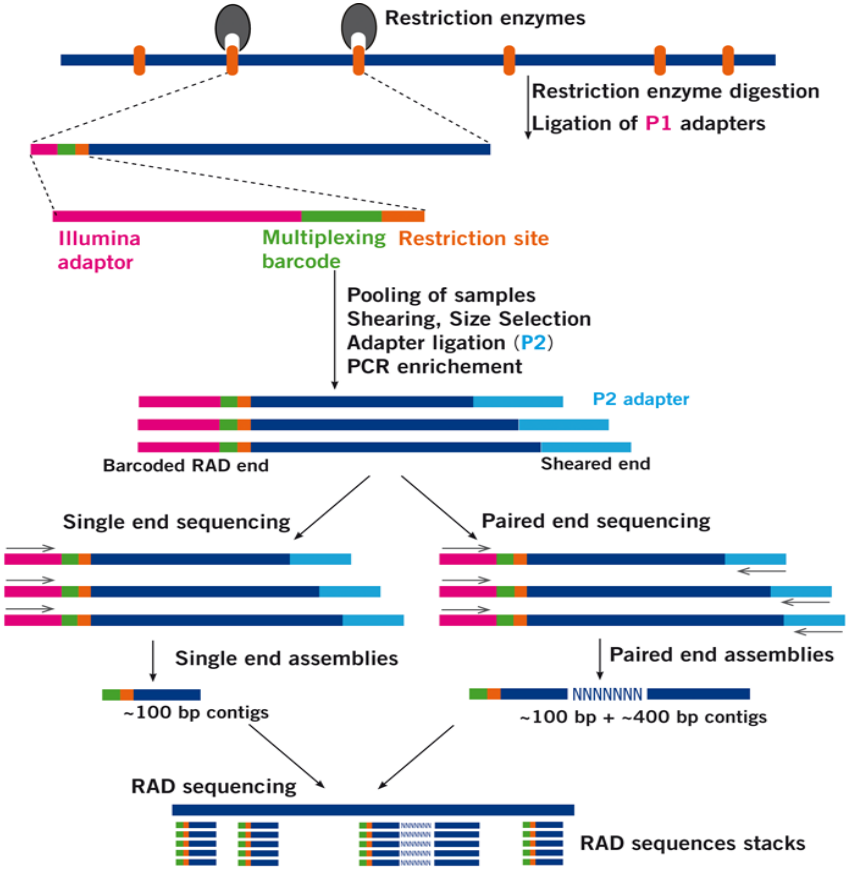


General – Restriction-site Associated DNA (RAD)

OPEN ACCESS Freely available online PLoS one

Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers

Nathan A. Baird[1☯], Paul D. Etter[1☯], Tressa S. Atwood[2], Mark C. Currey[3], Anthony L. Shiver[1], Zachary A. Lewis[1], Eric U. Selker[1], William A. Cresko[3], Eric A. Johnson[1*]



Pros

- Low cost; Degree of genome reduction is manipulatable based on RE selected
- Gathering large number of anonymous markers from large number of samples over a short period of time
- Widely used for inferring population structures, phylogeography, trait mapping, genetic maps, and association – plenty of case studies
- Bioinformatic pipelines (Stacks, pyRAD, etc.) are well established with good community support

Cons

- Need high quality DNA; not suitable for museum specimens
- Many steps in the workflow
- Difficult for large, highly repetitively genomes
- Problem with locus drop-outs/variance of depth across loci & individuals
- Phylogenetics: homology may be difficult to obtain between distantly related taxa due to mutations at cleavage sites
- Not the best choice of detecting selection when reference genome resource is lacking

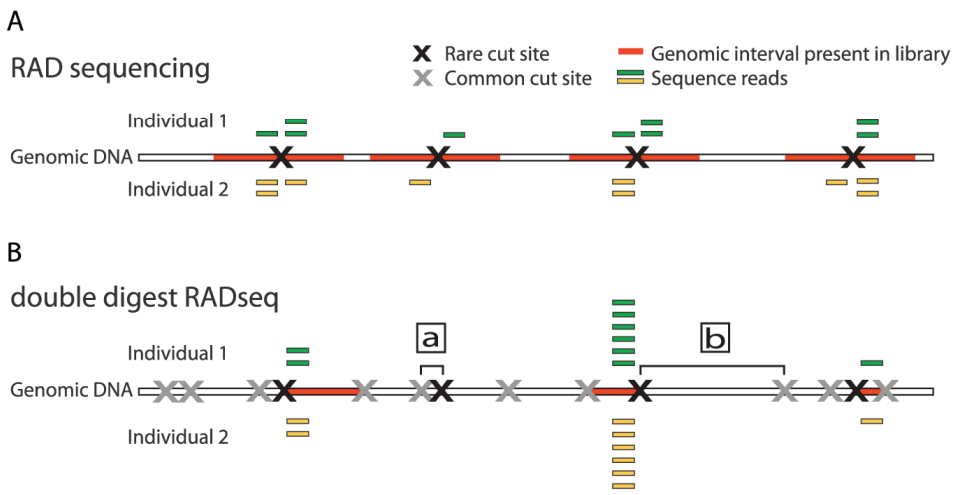
General – Double Digest RADseq (ddRAD)

OPEN ACCESS Freely available online

PLoS one

Double Digest RADseq: An Inexpensive Method for *De Novo* SNP Discovery and Genotyping in Model and Non-Model Species

Brant K. Peterson*, Jesse N. Weber, Emily H. Kay, Heidi S. Fisher, Hopi E. Hoekstra



Pros

- Compare to single digest RAD, ddRAD targets on sequencing fewer number of random markers to get a deeper coverage from large number of population samples
- Easier protocol without a shearing step during the library prep
- More flexible and control over the fragments that are sequenced

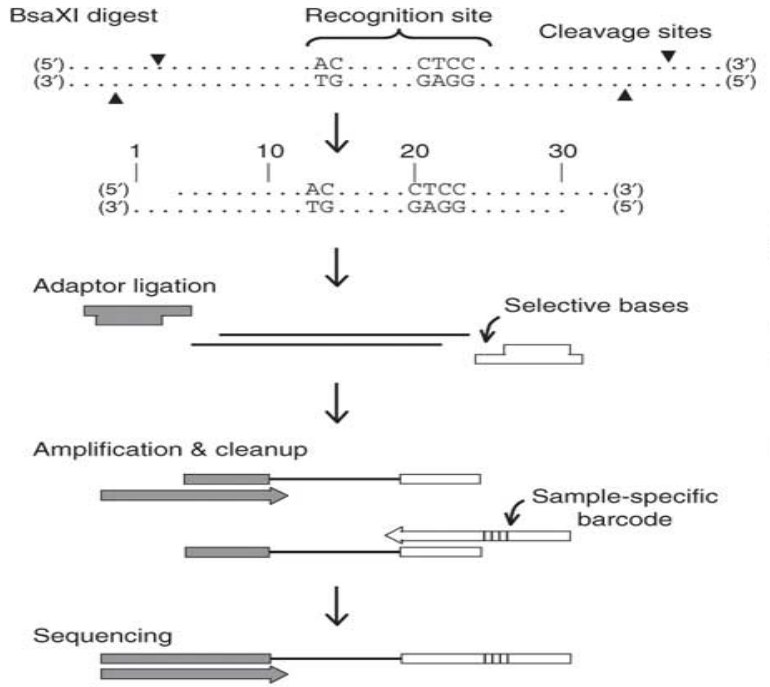
Cons

- Impossible to remove PCR duplications
- Size selection using a Pinpin is desired
- Use a frequent cutter: most likely to gain & lose fragments that results in large amount of missing data
- Also have similar issues in single RAD (genome size, selection, phylogenetics)

General – RAD by Type IIB Restriction Enzymes (2b-RAD)

2b-RAD: a simple and flexible method for genome-wide genotyping

Shi Wang[1,2,5], Eli Meyer[1,3,5], John K McKay[4] & Mikhail V Matz[1] Nature Methods 9, 808–810 (2012)



Type IIB enzymes (for example, BsaXI and AlfI) cleave genomic DNA upstream and downstream of the target site, producing tags of uniform length

Pros

- Easier workflow (compared to RAD)
- Very cost-effective
- Suitable for rapid, parallel genotyping of large numbers of samples with small genomes – "sequencing markers with higher throughput"

Cons

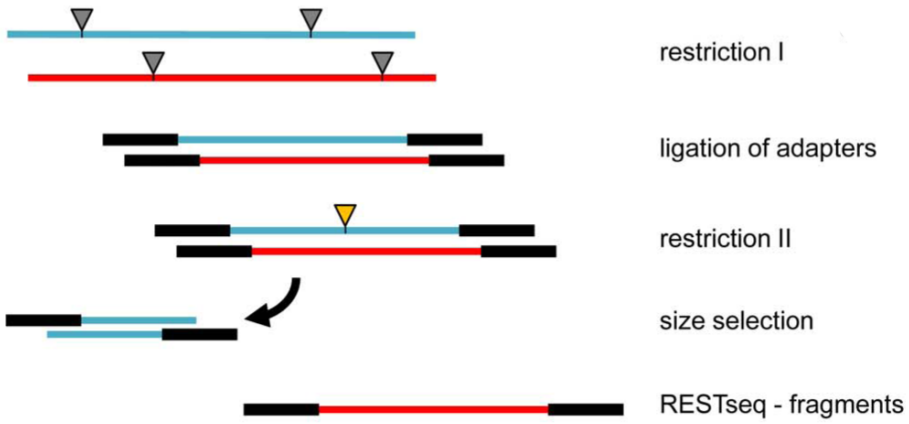
- Not widely used nowadays because tags are too short (e.g. 21–36 bp)
- Alignment is difficult

General – Restriction Fragment Sequencing (RESTseq)

OPEN ACCESS Freely available online PLOS ONE

RESTseq – Efficient Benchtop Population Genomics with RESTriction Fragment SEQuencing

Eckart Stolle[1,2*], Robin F. A. Moritz[1,3,4]



The method starts with the digestion of genomic DNA with a frequent cutting restriction endonuclease (e.g. TaqI (T, T^CGA)) to generate a high number of fragments. Following the ligation of platform-specific sequencing adapters, a second digestion with one or more frequent cutting restriction endonuclease (e.g. ApoI/MseI/BstUI) is further reducing the library in complexity.

Pros

- Compared to RAD/ddRAD, it has the advantage of a direct control over the complexity of the library due to the digestion after adapter ligation as well as the more unbiased fragment distribution
- It is robust for genomes with low complexities (e.g. reduce number of AT-rich fragments by using MseI(TruI) (TTAA) as the second RE)
- It allows for marker enriched SNP typing on small scale platforms and can increase the efficiency for analyzing large numbers of barcoded samples in high-throughput systems

Cons

- Share most issues seen in other RAD variants

General – RAD with Standard Illumina Library Preparation (ezRAD)

ezRAD: a simplified method for genomic genotyping in non-model organisms

PeerJ

Robert J. Toonen[1], Jonathan B. Puritz[1], Zac H. Forsman[1], Jonathan L. Whitney[1], Iria Fernandez-Silva[1], Kimberly R. Andrews[1] and Christopher E. Bird[2]

ezRAD differs from other RAD methods primarily through its use of standard Illumina TruSeq library preparation kits, which makes it possible for any laboratory to send out to a commercial genomic core facility for library preparation and next-generation sequencing with virtually no additional investment beyond the cost of the service itself.

Pros

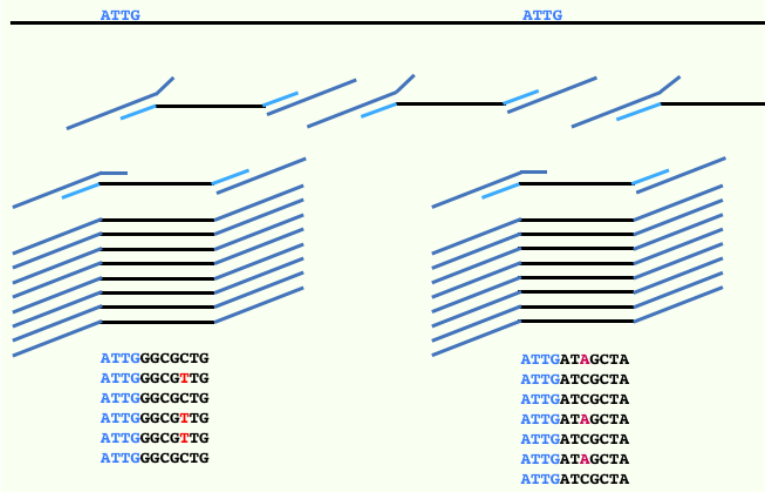
- Convenient: DNA samples can be sent to a sequencing facility for library prep and sequencing

Cons

- The sequencing core facility will charge a lot for library preparation
- Share common issues seen in other RAD variants

General – nextRAD (commercial)

SNPsaurus
GENOMES to GENOTYPES



Selective primers productively amplify only fragments with the selective sequence at the end. Subsequent sequencing of the library allows determination of genetic variation at these loci.

Pros

- Very few steps in the protocol: input DNA as low as 1-5ng!
- No need for size selection (selective primer side plus randomly-sheared side)
- Without the use of frequent-cutting REs
- Entire reads can be genotyped. In RAD the cutting cites have to be sequenced

Cons

- NextRAD can't use methylation-sensitive restriction enzymes, so have to pick a selective primer carefully to avoid amplifying repeats (but successful in plants with a >2 Gb genome).

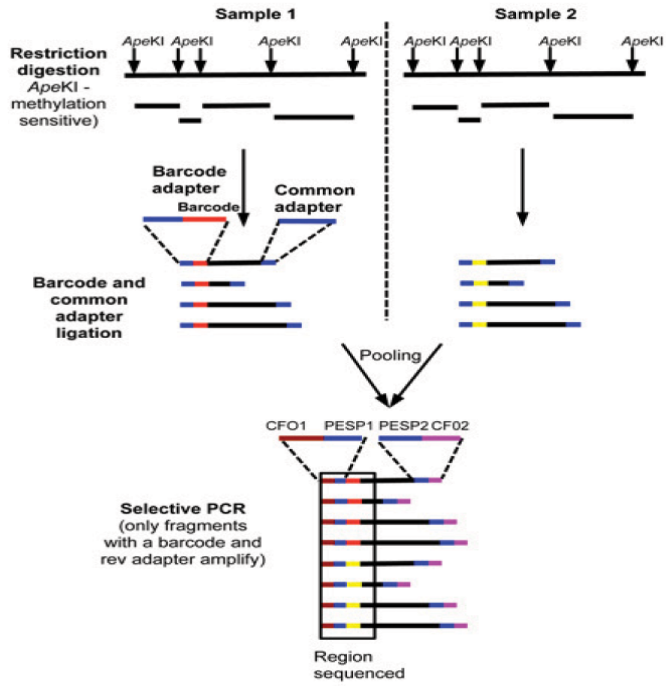
(information about pros and cons is kindly provided by Eric Johnson, inventor of RAD and nextRAD)
For more information about nextRAD please go to: http://snpsaurus.com/nextrad-genotyping/

General – Genotyping-By-Sequencing (GBS)

OPEN ACCESS Freely available online PLoS one

A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species

Robert J. Elshire[1], Jeffrey C. Glaubitz[1], Qi Sun[2], Jesse A. Poland[3], Ken Kawamoto[1], Edward S. Buckler[1,4], Sharon E. Mitchell[1]*



Most GBS relies on the use of methylation-sensitive restriction endonucleases (e.g., ApeKI for plants) to avoid repetitive regions of the genome, while targeting lower copy regions of the genome (http://www.igd.cornell.edu/index.cfm/page/GBS/gbsfaq.htm)

Pros

- Compared to RAD, it has more simplified protocol, requiring less DNA, no sonication and size selection
- More widely used in plants with large genomes (e.g. barley, soybeans, maize, switchgrass, wheat, rice, grape, cacao etc.)
- Designed for lightly sequencing large number of markers and imputing the missing genotypes from the many reference genomes available
- Bioinformatics tools (e.g. TASSEL, UNEAK)

Cons

- Although with different adapters/barcodes and general workflows of library preparation, it is essentially very similar to RAD -> it still shares most cons with RAD
- Light sequencing of many loci -> low coverage & more missing data -> need a solid reference for alignment to infer nearby genotypes

General – Double-digest Genotyping-By-Sequencing (ddGBS)

OPEN ACCESS Freely available online

PLoS one

Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach

Jesse A. Poland[1,2*], Patrick J. Brown[3], Mark E. Sorrells[4], Jean-Luc Jannink[4,5]

Two enzyme digestion, forward barcoded adapter and reverse, common, Y-adapter

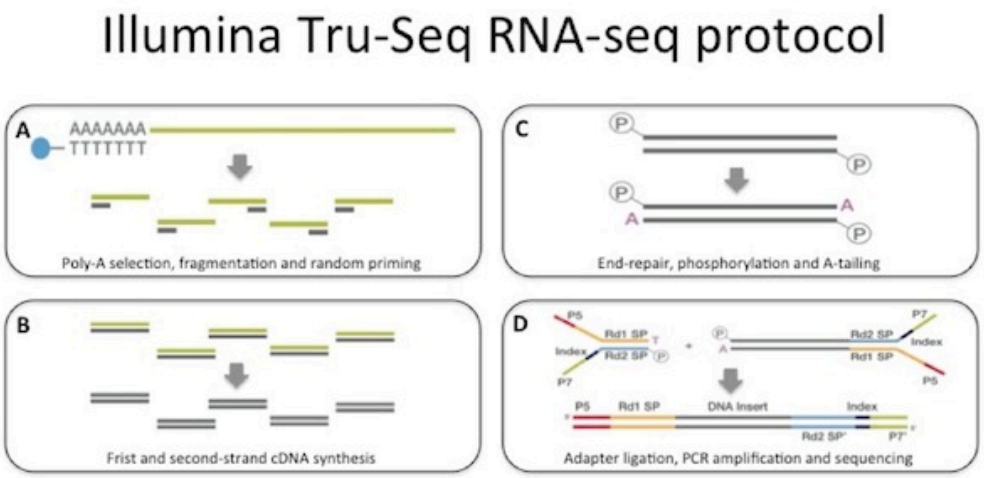
Pros

- Share pros with single-digest GBS

Cons

- Share cons with single-digest GBS

General – Transcriptome Sequencing (RNAseq for Marker Development)



Library prep begins from 100ng-1ug of Total RNA which is poly-A selected (A) with magnetic beads. Double-stranded cDNA (B) is phosphorylated and A-tailed (C) ready for adapter ligation. The library is PCR amplified (D) ready for clustering and sequencing.

Pros

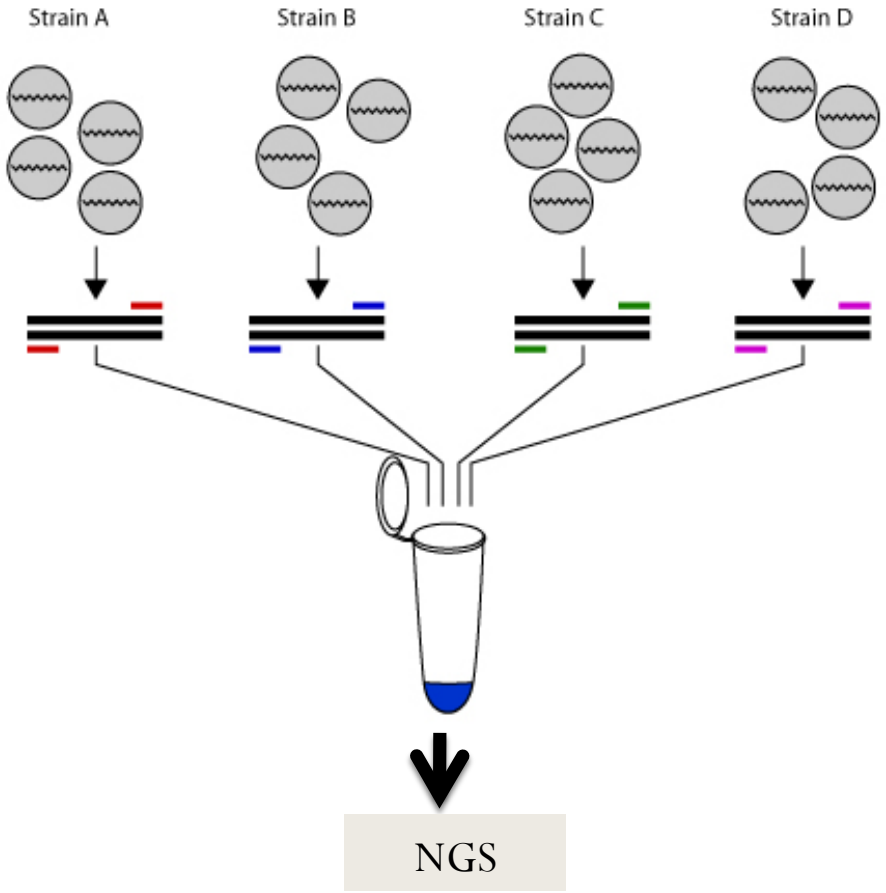
- Targeting sequences and relative abundance of transcribed portion of the genomes
- Easily generate thousands to over ten thousands of markers from multiple samples
- Can serve as reference for marker selection for exon captures

Cons

- Require RNA samples
- Alleles may vary in transcription level across genes, tissues and stages, and thus may generate inaccurate genotyping data
- Library preparation expensive
- Highly expressed genes dominate sequence data
- Not cost-effective for large scale phylogenomic or population genomic applications

Target Enrichment – Multiplexed Amplicon Sequencing

The simplest version of amplicon sequencing



Pros

- Metagenomics applications (e.g. 16S for investigating diversity and structure of complex microbial communities & populations)
- Cost-effective for targeting a small number of loci from large number of samples
- For very large genomes (eg. >100Gb), amplicon sequencing may be the only/best option

Cons

- Primer design and optimizing PCR
- For homemade protocol: difficulty of scaling PCR to the capacity of NGS platforms - very labor intensive / costly beyond a few loci

Target Enrichment – Primer Extension Capture (PEC)

17 JULY 2009 VOL 325 SCIENCE

Targeted Retrieval and Analysis of Five Neandertal mtDNA Genomes

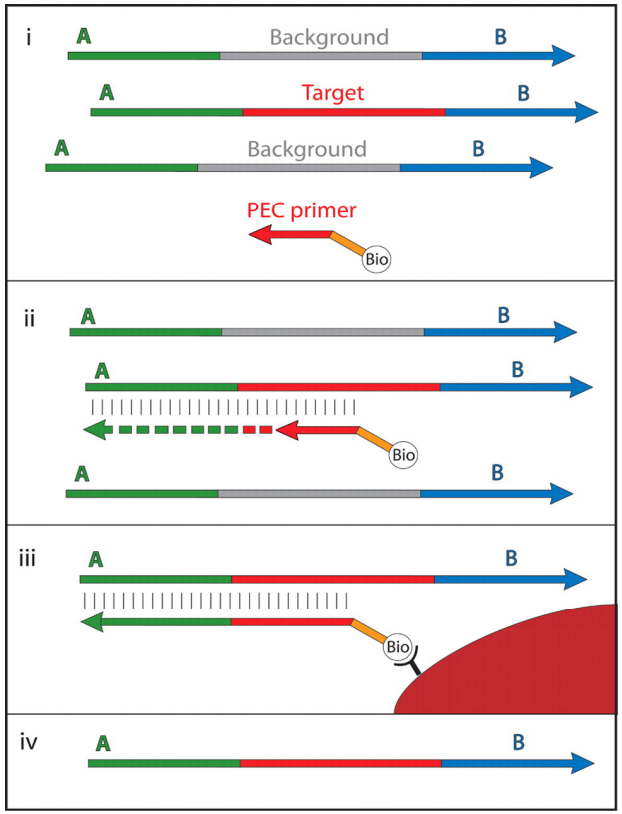
Adrian W. Briggs,[1*] Jeffrey M. Good,[1] Richard E. Green,[1] Johannes Krause,[1] Tomislav Maricic,[1] Udo Stenzel,[1] Carles Lalueza-Fox,[2] Pavao Rudan,[3] Dejana Brajković,[4] Željko Kućan,[3] Ivan Gušić,[3] Ralf Schmitz,[5,6] Vladimir B. Doronichev,[7] Liubov V. Golovanova,[7] Marco de la Rasilla,[8] Javier Fortea,[8] Antonio Rosas,[9] Svante Pääbo[1]

Anneal to biotinylated Primers

A single Taq DNA polymerase extension

Captured by streptavidin-coated magnetic beads

Elution



Pros

- The PEC procedure is particularly well suited for enrichment of smaller genomic regions (complete mtgenome and a few nuclear loci) from highly degraded DNA samples, as is commonly retrieved from historical or ancient specimens
- Require an initial investment in stocks of targeted probes that can then be used on large collections of samples

Cons

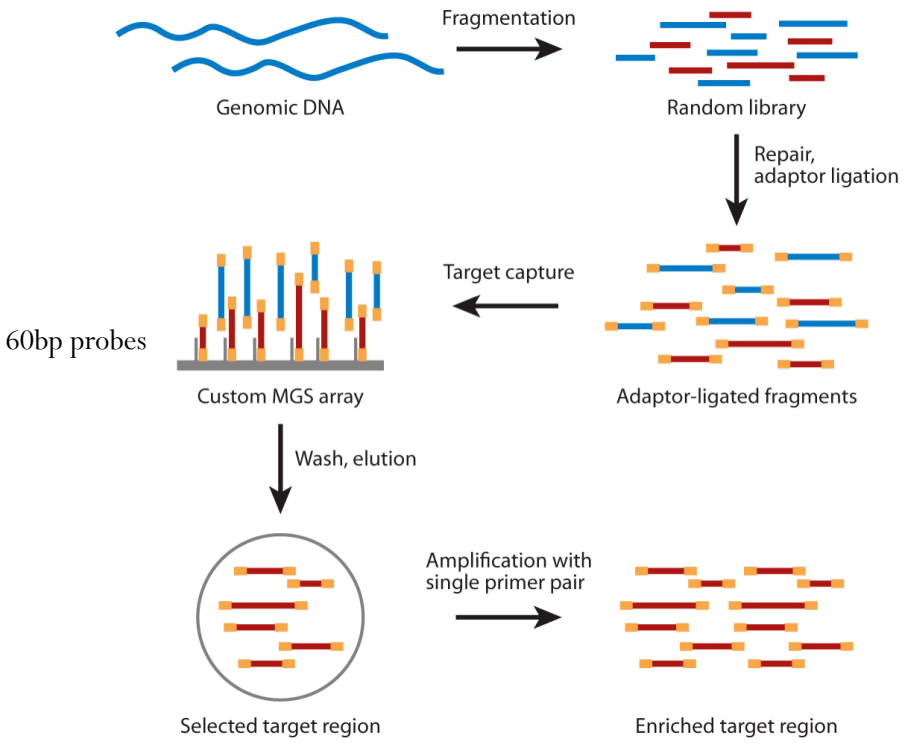
- Share issues with multiplexed amplicon sequencing
- Synthesis of many biotinylated primers is not cheap
- Not necessary for high quality DNA samples

Target Enrichment – Microarray-based Exon 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.



Agilent Custom SureSelect microarray 1M or 244K format

Pros

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

Cons

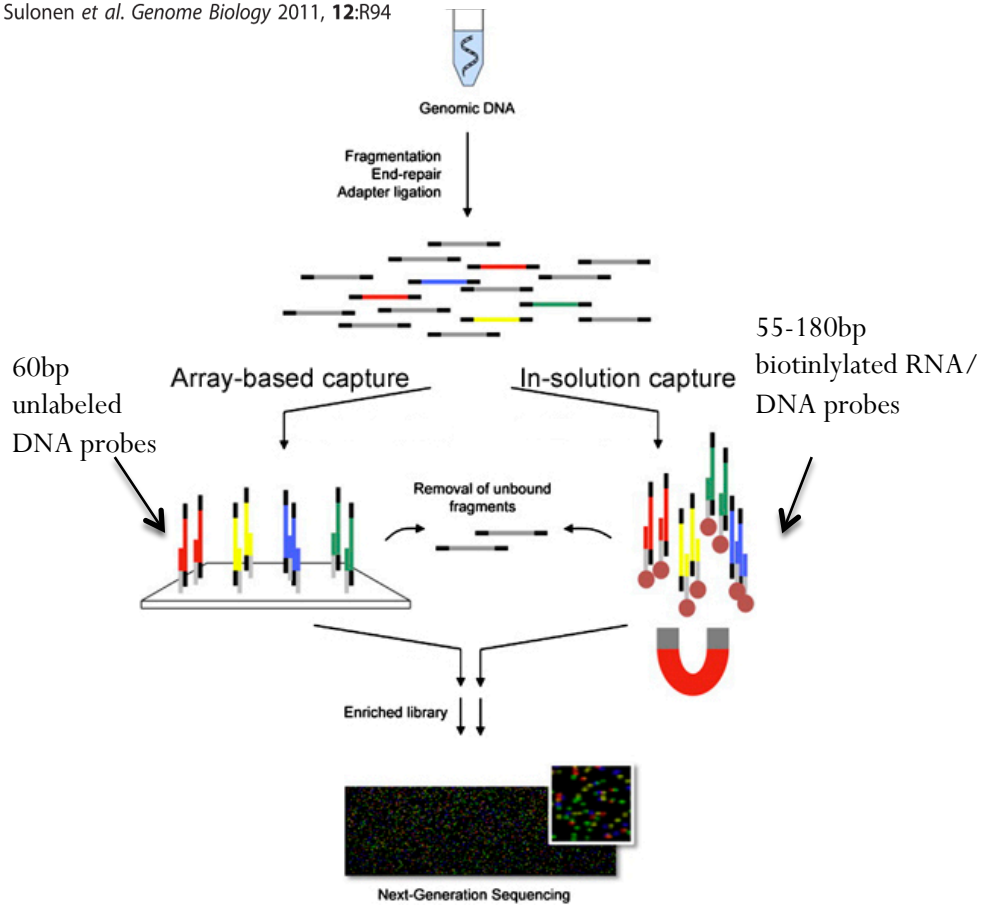
- Need a reference for probe design (also true for other types of hybridization-based methods.)
- Low capture efficiency
- Probe length short (60bp)
- Need special equipments (hybridization chamber, gasket slides, oven, etc.)
- Not cost-effective for surveying large number of samples
- Need large amount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

Target Enrichment - In-solution-based Exon Capture

Comparison of solution-based exome capture methods for next generation sequencing

Anna-Maija Sulonen[1,2], Pekka Ellonen[1], Henrikki Almusa[1], Maija Lepistö[1], Samuli Eldfors[1], Sari Hannula[1], Timo Miettinen[1], Henna Tyynismaa[3], Perttu Salo[1,2], Caroline Heckman[1], Heikki Joensuu[4], Taneli Raivio[5,6], Anu Suomalainen[3] and Janna Saarela[1*]

Sulonen *et al. Genome Biology* 2011, **12**:R94



For non-model systems:

- Agilent SureSelect XT/XT2 Custom kits
- NimbleGen SeqCap EZ Developer kits
- MYcroarray mybaits kits

Pros

- Target size large (e.g. up to 200 Mb for NimbleGen)
- Low amount of input DNA and Cot-1 DNA
- High level of multiplexing (>50)
- Suitable for large-scale population genomic projects (NimbleGen) or phylogenomic projects (MyBait)
- High capture efficiency
- Automation friendly (no special equipments needed)

Cons

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

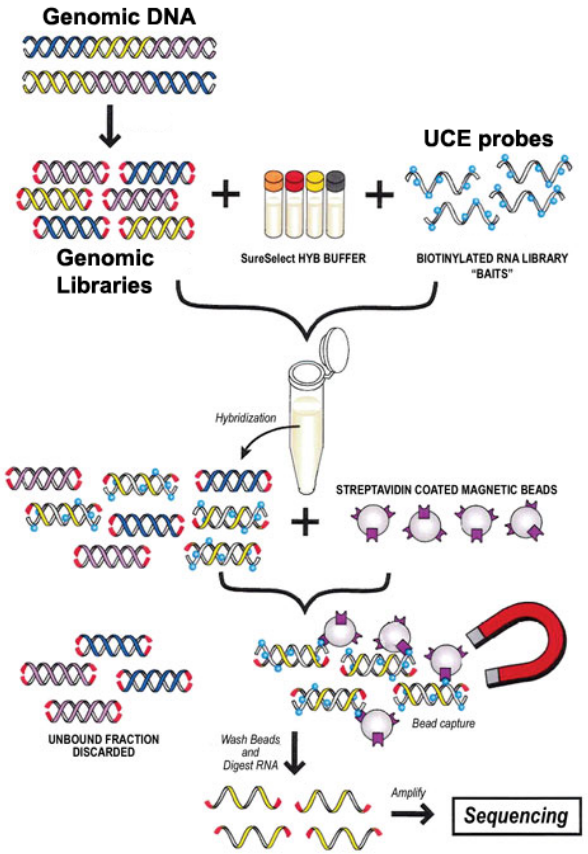
Target Enrichment - Ultraconserved Elements (UCEs) Capture

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)

Target Enrichment- Anchored Hybrid Enrichment (AHE)

Syst. Biol. 0(0):1–18, 2012

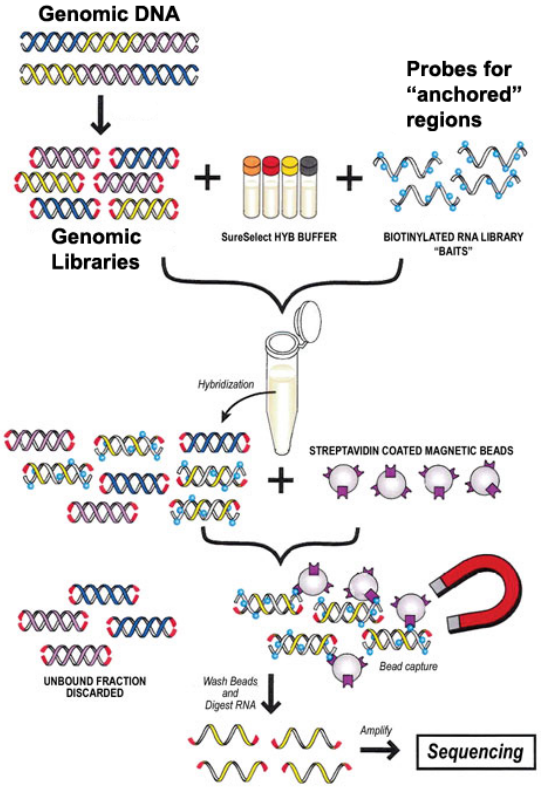
DOI:10.1093/sysbio/sys049
Advance Access publication on May 17, 2012

Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics

ALAN R. LEMMON[1,*], SANDRA A. EMME[2], AND EMILY MORIARTY LEMMON[2]

[1]*Department of Scientific Computing, Florida State University, Dirac Science Library, Tallahassee, FL 32306-4102, USA; and* [2]*Department of Biological Science, Florida State University, 319 Stadium Drive, PO Box 3064295, Tallahassee, FL, 32306-4295, USA;*
**Correspondence to be sent to: Department of Scientific Computing, Florida State University, Dirac Science Library, Tallahassee, FL 32306-4102; E-mail: alemmon@fsu.edu.*

Received 1 November 2011; reviews returned 19 January 2012; accepted 7 May 2012
Associate Editor: Bryan Carstens



Center for anchored phylogenomics

http://anchoredphylogeny.com/training/

Pros

- Target 500 loci across vertebrates (fewer loci for more samples with deeper coverage)
- Captured genomic fragments contain both conserved regions (but not ultra-conserved) and variable (flanking) regions so it is robust for resolving both deep- and shallow-phylogenies
- A more densely-tiled probes to represent several relevant lineages
- Collaboration (you send them DNA and you get trees back)

Cons

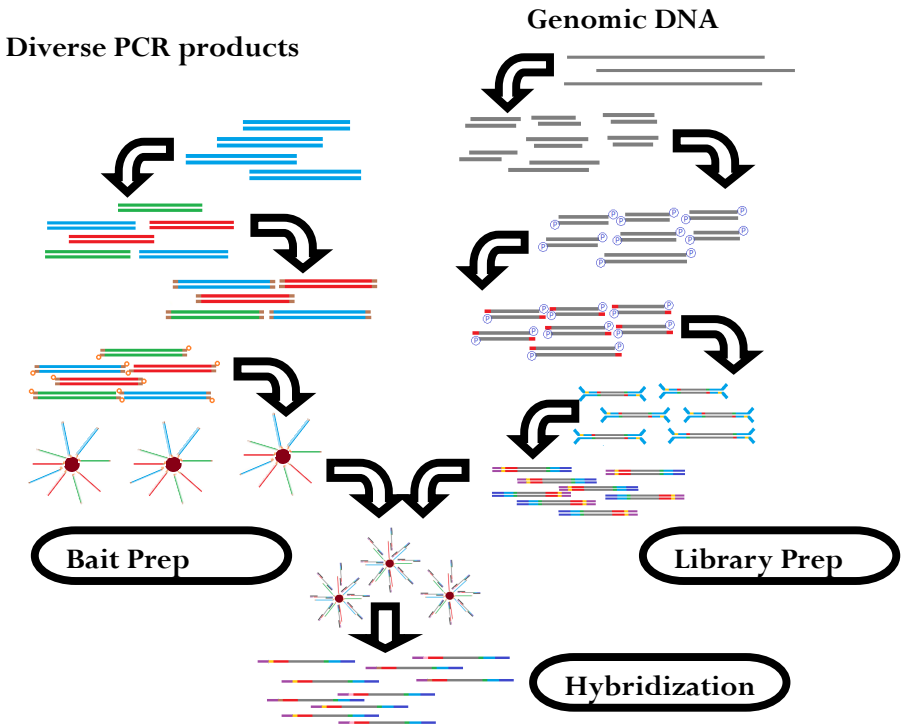
- Technical and bioinformatic pipelines not available (?)

Target Enrichment- Sequence Capture using PCR-generated Probes (SCPP) (home-made)

Sequence capture using PCR-generated probes: a cost-effective method of targeted high-throughput sequencing for nonmodel organisms

JOSHUA V. PEÑALBA,* LYDIA L. SMITH,*† MARIA A. TONIONE,*‡ CHODON SASS,§ SARAH M. HYKIN,* PHILLIP L. SKIPWITH,* JIMMY A. MCGUIRE,*† RAURI C. K. BOWIE*† and CRAIG MORITZ†¶

Molecular Ecology Resources (2014) 14, 1000–1010



This technique is developed by a group for MVZ researchers, mainly led by Josh Penalba, Lydia Smith et al.

Pros

- Low cost
- Very high enrichment efficiency
- Highly robust for both shallow and deep-level phylogenies (up to 200 loci with hundreds of samples)
- Biotinylated baits are easy to re-generate and once generated, they can be used for many projects

Cons

- Prior knowledge for the PCR loci (a lot of primer design)
- Labor-intensive for doing and optimizing a lot of independent PCR reactions

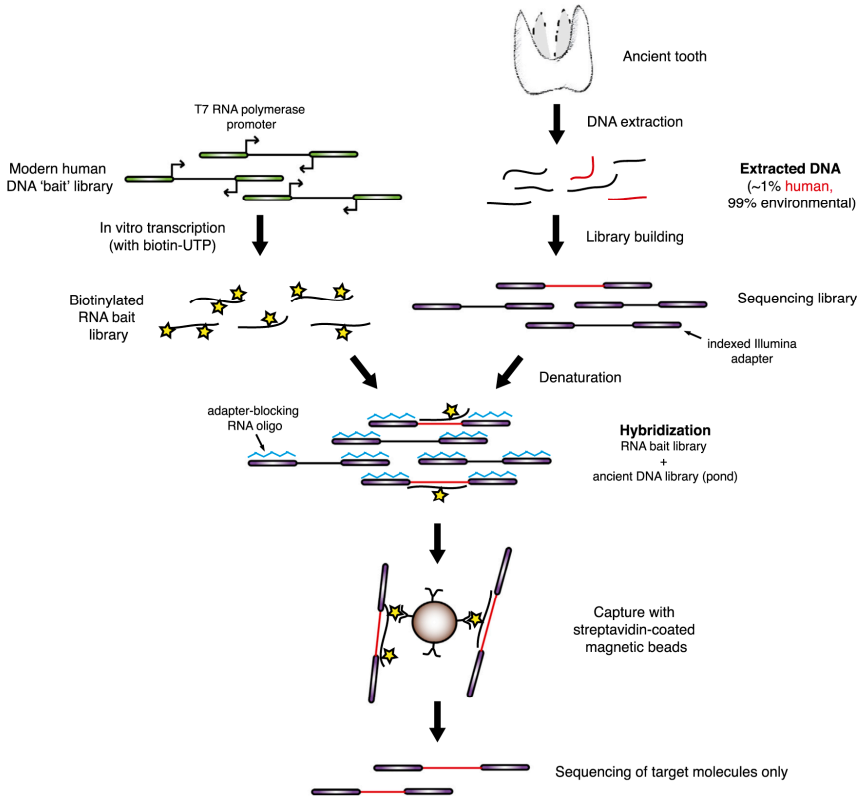
Target Enrichment - Whole Genome In-Solution Capture

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,*]



In-house or commercial (Mybaits)

Pros

- Work effectively for enriching genomic DNA from ancient specimens that contain very low levels of endogenous DNA (<1%)
- The depleted (environmental) DNA can be sequenced for metagenomics analyses.

Cons

- Not necessary for modern DNA samples
- Not feasible for organisms with extremely low complexity genomes and/or large genomes

Other types of genome reduce representation methods

- Reduced-Representation libraries (RRLs)
- Complexity Reduction of Polymorphic Sequences (CRoPS)
- Genome Reducing and Sequencing (GGRS)
- Molecular Inversion Probes (MIPs)
- Connector Inversion Probe (CIPs)
- SNP genotyping

Too many options?

Visit us at the CGRL and we can provide consultations for your project design

Choosing a right approach for your project .....

- **It depends on your questions:**

• *Population genetic parameter estimate (thetas, structure, hybridization, gene flow, change in Ne, etc.)*: RAD, GBS, exon capture

• *Selection on protein evolution*: RNAseq, exon capture

• *Both demography and selection*: exon capture (exons + introns)

• *Other population genetic applications (admixture mapping, constructing linkage map, associations, phylogeography, etc.)*: RAD, GBS, exon capture

• *Larger number (>10,000) of phylogenetic markers*: RNAseq, exon capture, (RAD if shallow divergence)

• *Small number (100s) of phylogenetic markers*: Amplicon sequencing, AHE, SCPP, UCEs

- It depends on the quality and quantity of the DNA

• *Low quality (heavily degraded) DNA*: All hybridization-based methods

• *Low quantity (e.g. a few nanograms)*: nextRAD; certain library prep protocols for sequence captures/RAD

- **It depends on the desired sample size (S) + target size (T)**

• *Large S + large T*: exon capture, RAD, GBS

• *Large S + small T*: Amplicon sequencing, AHE, UCE, PEC, SCPP, RESTseq

- **It depends on your budegt**

• *Expensive*: All commercial in-solution exon capture kits

• *Cheap*: All RAD/GBS variants, SCPP, array-based exon capture, UCEs etc.

- **It also depends on the genome size/composition, availability of reference resources, the timeline for getting the project finished, experience and support in lab and bioinformatics……**

A General Workflow for *de novo* Exon Capture

Bi *et al. BMC Genomics* 2012, **13**:403
http://www.biomedcentral.com/1471-2164/13/403

BMC Genomics

METHODOLOGY ARTICLE Open Access

Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales

Ke Bi[1*], Dan Vanderpool[2], Sonal Singhal[1,3], Tyler Linderoth[1,3], Craig Moritz[1,3] and Jeffrey M Good[2]

MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 6018–6032 doi: 10.1111/mec.12516

Unlocking the vault: next-generation museum population genomics

KE BI,* TYLER LINDEROTH,*† DAN VANDERPOOL,‡ JEFFREY M. GOOD,‡ RASMUS NIELSEN† and CRAIG MORITZ*†§

Museum of Vertebrate Zoology, University of California, 3101 Valley Life Sciences Building, Berkeley, California 94720, USA, †Department of Integrative Biology, University of California, 3060 Valley Life Sciences Building, Berkeley, California 94720, USA, ‡Division of Biological Sciences, University of Montana, Missoula, Montana 59812, USA, §Research School of Biology and Centre for Biodiversity Analysis, Australian National University, Canberra, ACT 0200, Australia

A General Workflow for *de novo* Exon Capture

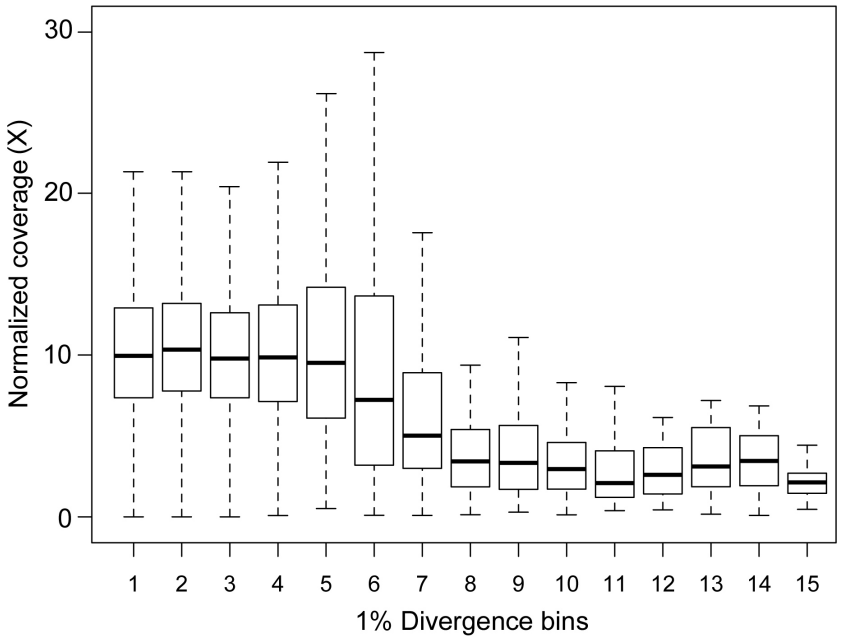
1. **Target selection and probe design**
2. **Library preparation and multiplexing**
3. **Exon capture experiments**
4. **Post-capture quality control**

Genomic Resources

"Non-model organisms": no pre-existing, closely related genome resources

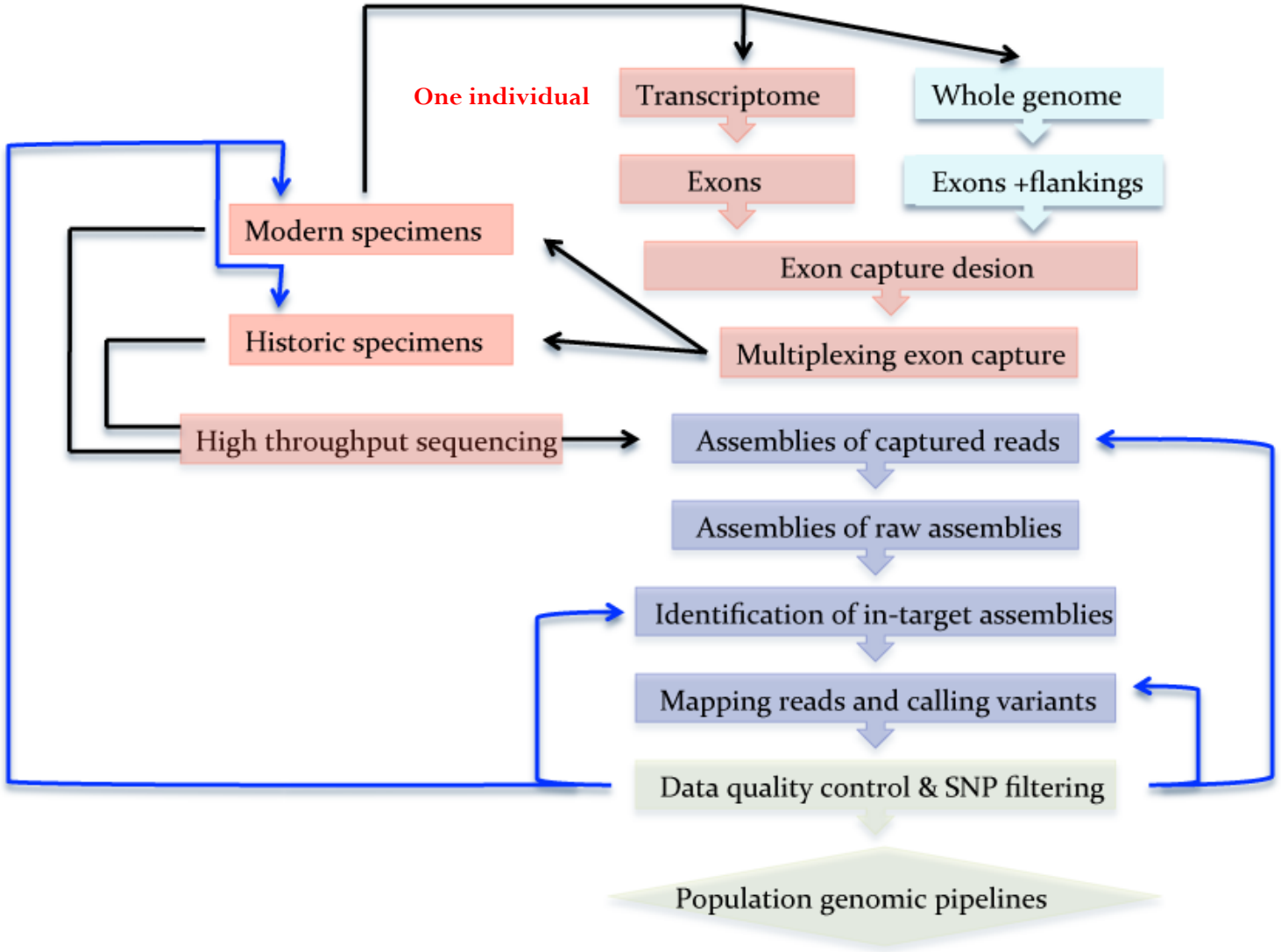
Designing exon captures using divergent reference can reduce enrichment efficiency

Capture efficiency vs. sequence divergence

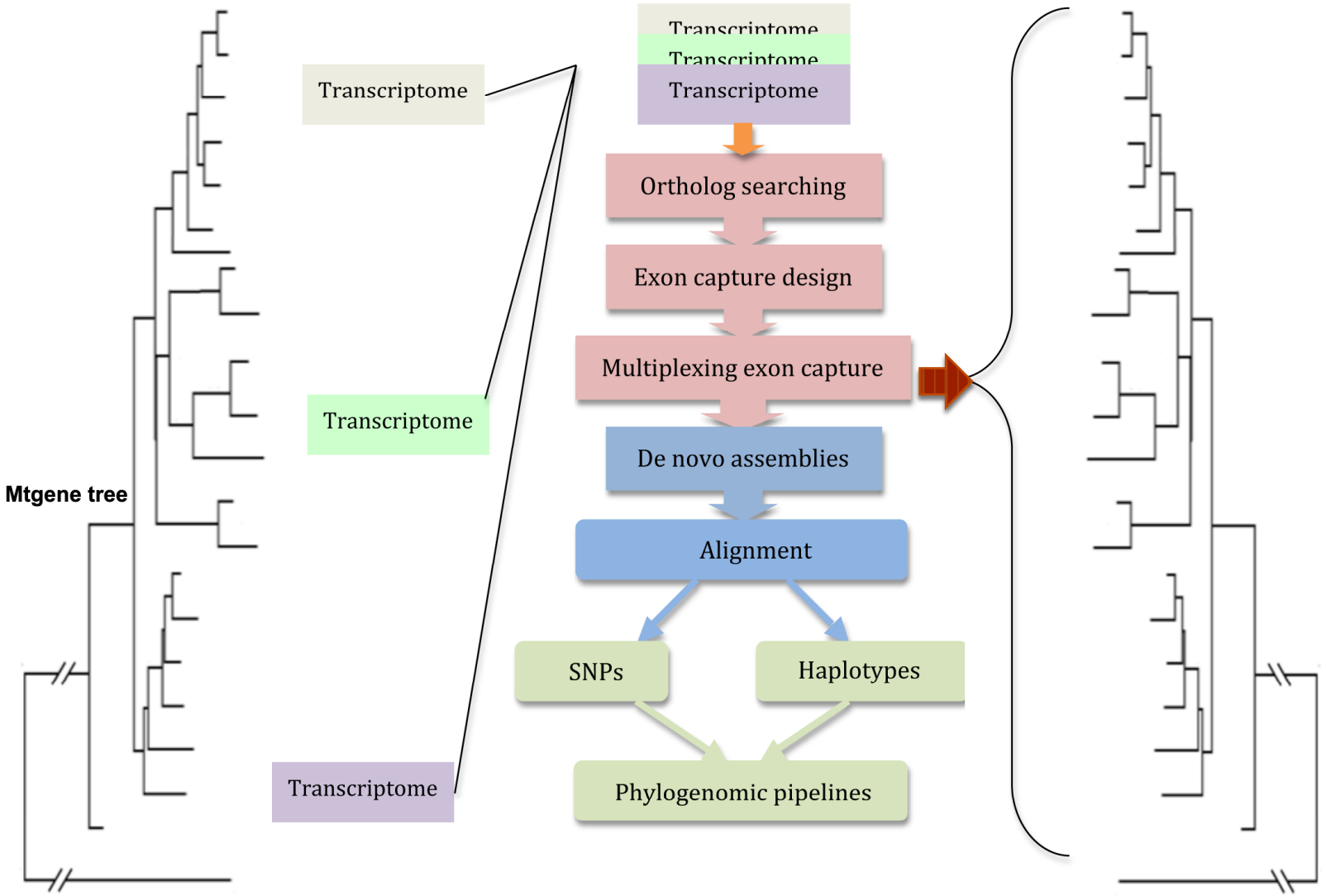


The level of coverage starts to decrease rapidly when the divergence becomes greater than 5%. Bi et al. 2012. BMC Genomics

Develop *de novo* Genomic Resources for Population Genomic Projects



Develop *de novo* Genomic Resources for Phylogenomic Projects

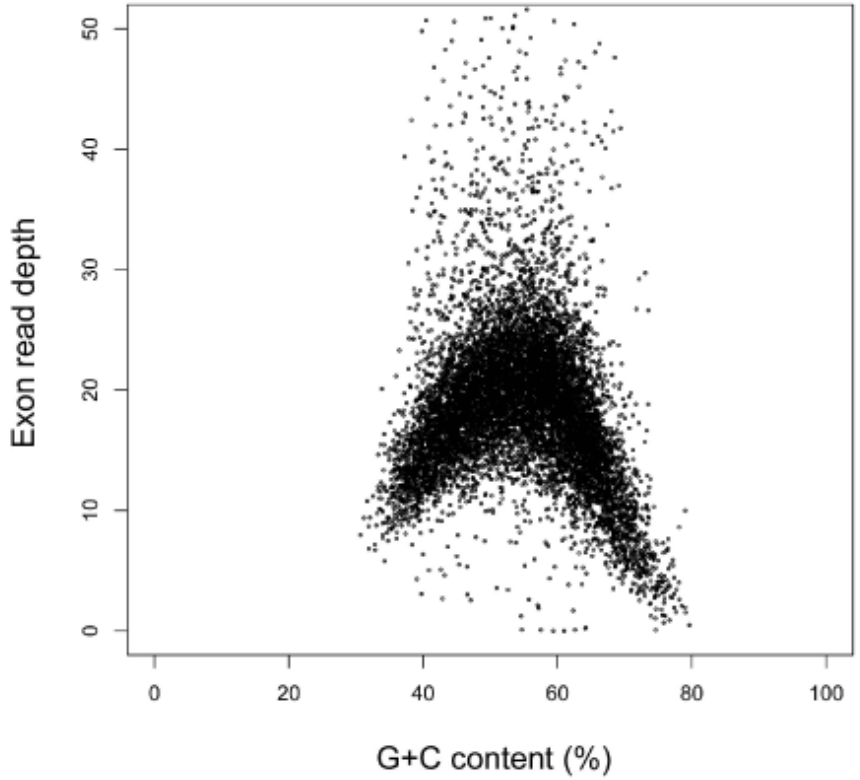


Initial Filtering on Candidate Target Loci

- Repeat masking (hard & soft)
- Remove regions with extreme G/C content
- Unique markers in target sets
- Length cutoff
- Positive (targeted loci) and negative (No-target) controls for qPCR assessment
- A **short** mitochondrial locus

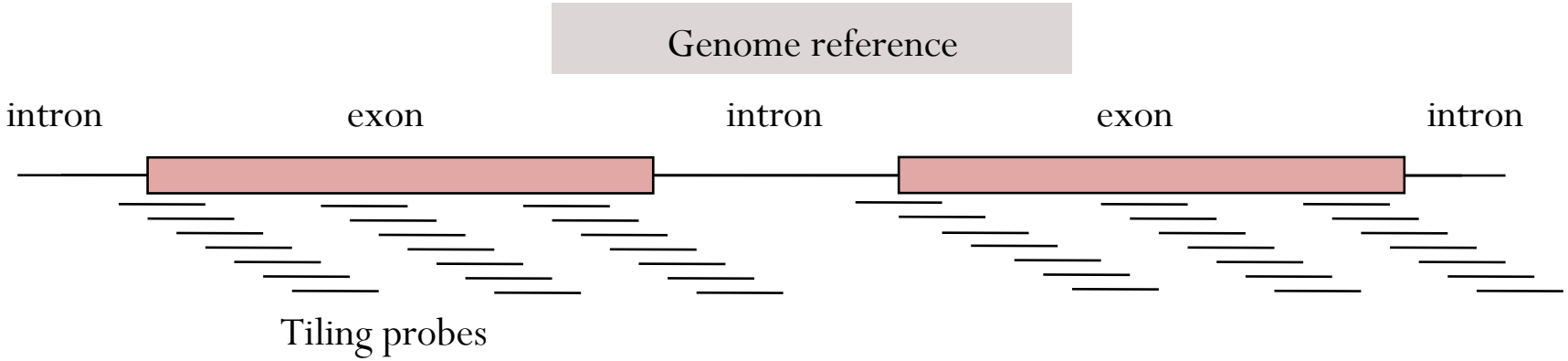
for evaluating empirical error rates

DO NOT include a long mt locus unless it is of particular interest!



Bi et al. 2012. BMC Genomics

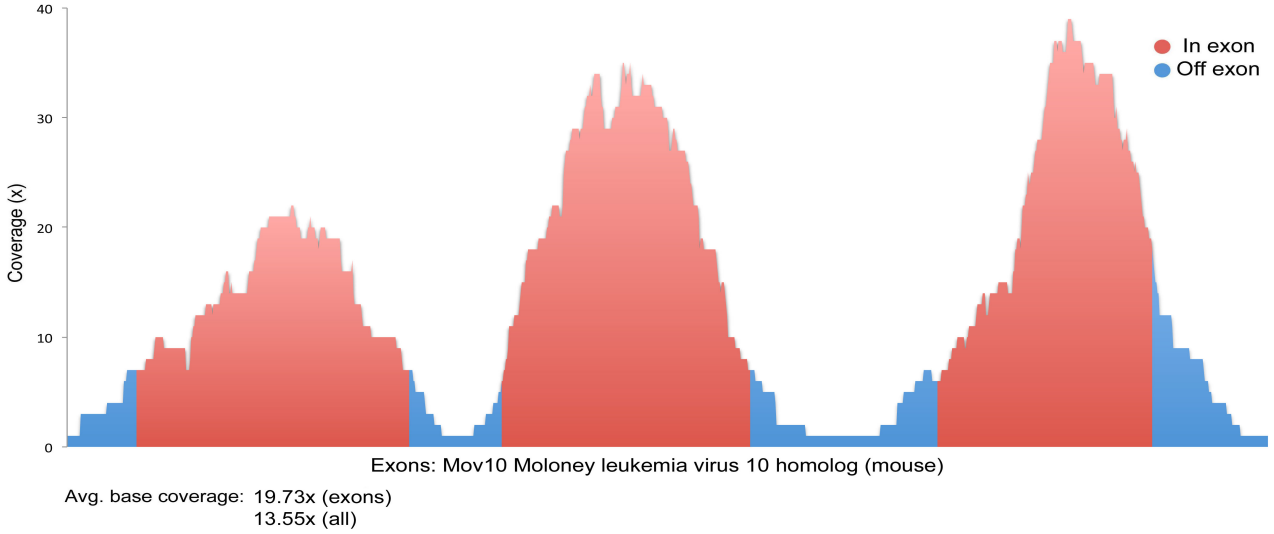
Probe Design from One Reference (Population Genomic Projects)

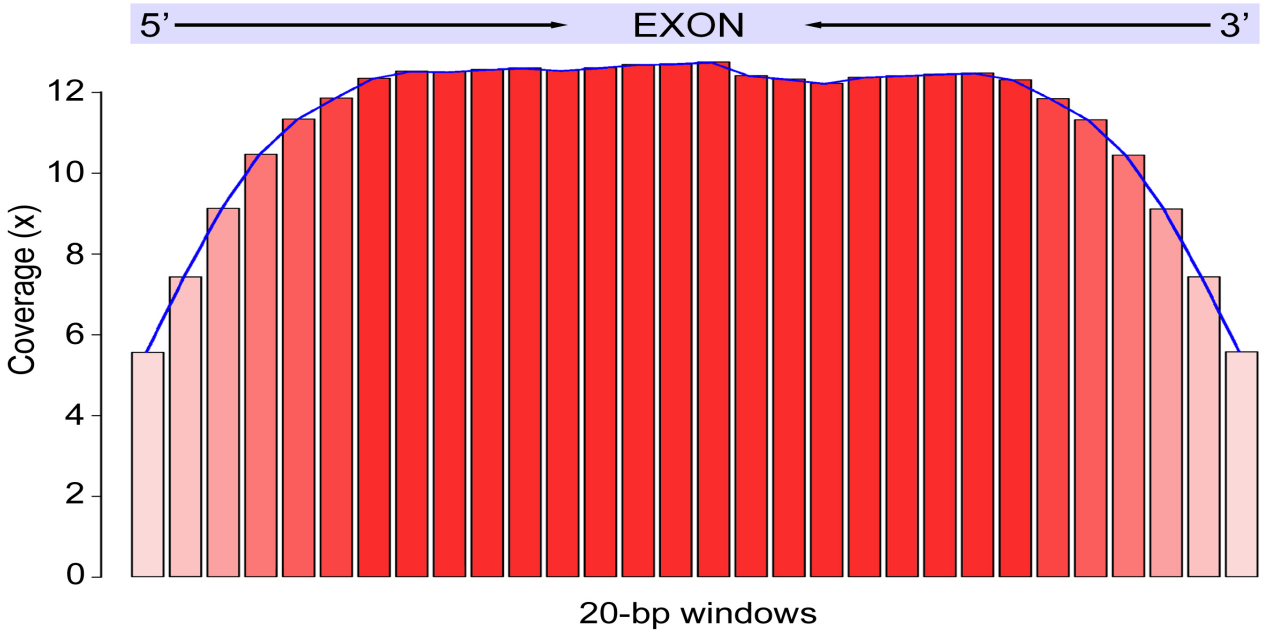


Probe Design from Multiple References (Phylogenomic Projects)

Exon capture with pooled, divergent baits to maximize enrichment of genetically divergent DNA libraries

Sequence Capture Edge Effect





In order to improve the coverage at the edge of exons and adjacent flanking introns or UTRs: probes can be more densely tiled at the edges of each exonic locus

A General Workflow for *de novo* Exon Capture

1. Target selection and probe design

2. **Library preparation and multiplexing**

3. Exon capture experiments

4. Post-capture quality control

Custom Genomic Library Preparation

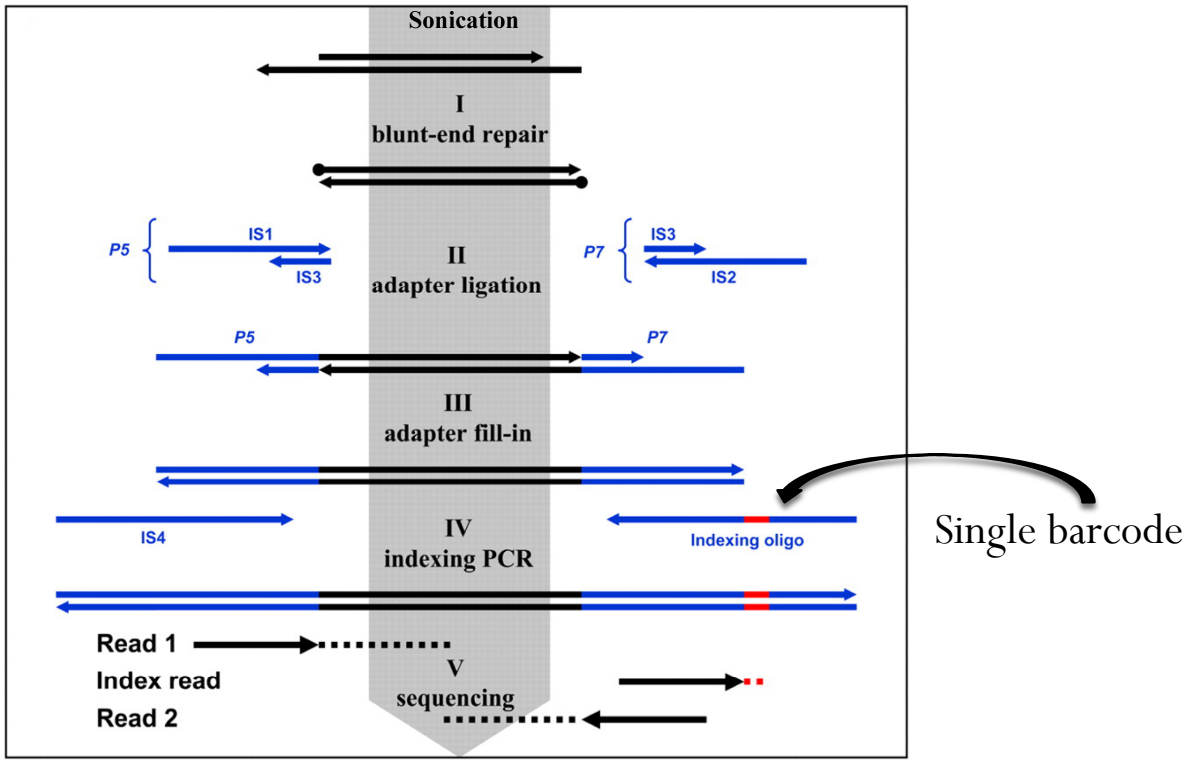
Protocol

Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing

Matthias Meyer[1] and Martin Kircher

Max Planck Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany

Cold Spring Harb Protoc, 2010 pdb.prot5448.

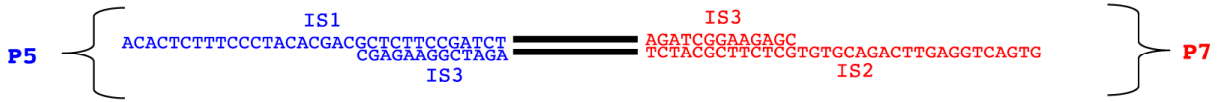


Nucleic Acids Research, 2011, 1–8
doi:10.1093/nar/gkr771

Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform

Martin Kircher*, Susanna Sawyer and Matthias Meyer*

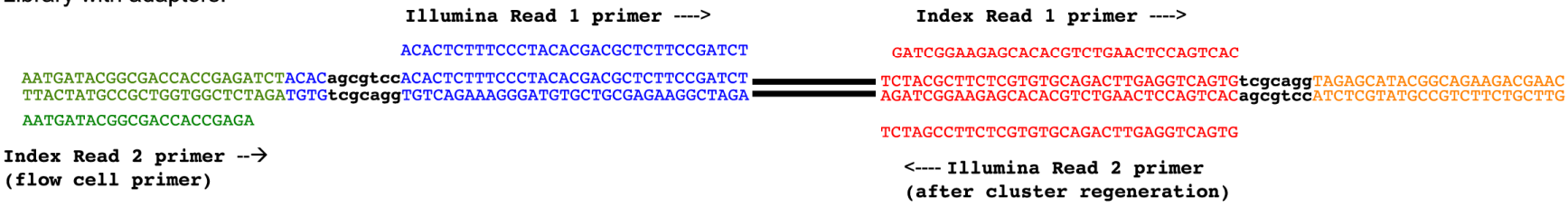
Adapter ligation:



Adapter fill-in:

Indexing PCR:

Library with adapters:



Per library cost (SeraMag)

			current egl price	volume	needed for a single indexed library	cost for a single indexed library		
SeraMag beads (in PEG solution)	Thermo Scientific		$1.00	mL	450µL	$0.51		all steps
ATP (100 mM)	Fermentas (Thermo Fisher)	FER R0441	$0.16	µL	0.7 µL	$0.13		Blunt-End Repair
Bst DNA polymerase, large fragment	New England BioLabs	M0275L	$0.21	µL	1.5 µL	$0.36		Adapter Fill-In
dNTPs	EGL stock	25 mM each	$6.90	100µL	1.48 µL	$0.12		Blunt-End Repair, Adapter Fill-In, Indexing PCR
Illumina adapters	IDT		$1.00	µL	1.0 µL	$1.13		Adapter Ligation
Indexing oligos	Sigma		$0.50	µL	1.0 µL	$1.13		Indexing PCR
Phusion Hot Start High-Fidelity DNA Polymerase 500U	Finnzymes (Thermo Fisher)	F-540L	$2.21	µL	1.0 µL	$2.49		Indexing PCR
Primer IS4	IDT		$0.03	µL	2.0 µL	$0.07		Indexing PCR
T4 DNA ligase	Fermentas (Thermo Fisher)	FER EL0012	$0.73	µL	1.0 µL	$0.83		Adapter Ligation
T4 DNA polymerase	Fermentas (Thermo Fisher)	FER EP0062	$1.90	µL	1.4 µL	$3.00		Blunt-End Repair
T4 polynucleotide kinase	Fermentas (Thermo Fisher)	FER EK0032	$0.76	µL	3.5 µL	$3.00		Blunt-End Repair
Tango buffer 10x	Fermentas (Thermo Fisher)	FER BY5	$2.90	mL	7.0 µL	$0.01		Blunt-End Repair
						$12.78		**reagent and bead costs**
								average of 2 indexing PCR reactions per library
								[includes EGL overhead]

Note: All costs given here and in all other slides are inclusive of sales tax, shipping, EGL overhead, etc. and provide the final researcher cost for that item.

Note: Recent trials suggest that the enzymes for the blunt ending reaction can be halved or even quartered without any ill-effect on the subsequent adapter ligation. This would drop reagent costs below $10/library.

Provided by the EGL manager
Lydia Smith

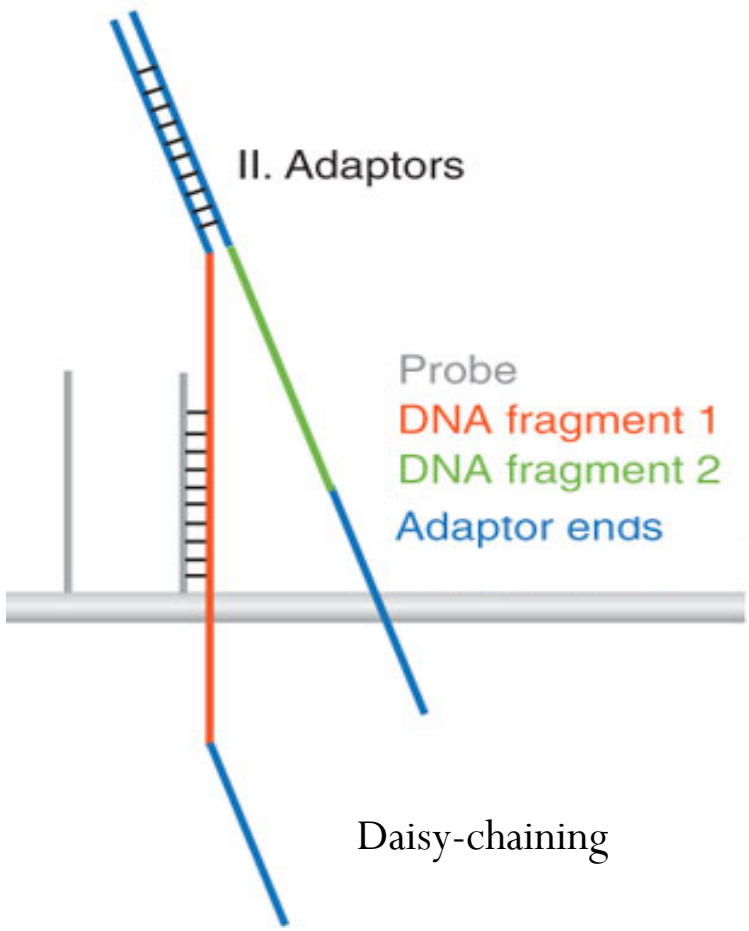
Meyer & Kircher Library Preparation Protocol

Pros

- Cost effective
- Robust for multiplexing
- Low input DNA (200 ~ 500 ng)

Cons

- The ligation of regular (long) Illumina adapters prior to exon capture that are more prone to binding DNA fragments that do not belong to the target region (daisy chaining) during hybridization
- Barcode blocking oligos (HPLC) that are needed to block each barcode adapter are expensive



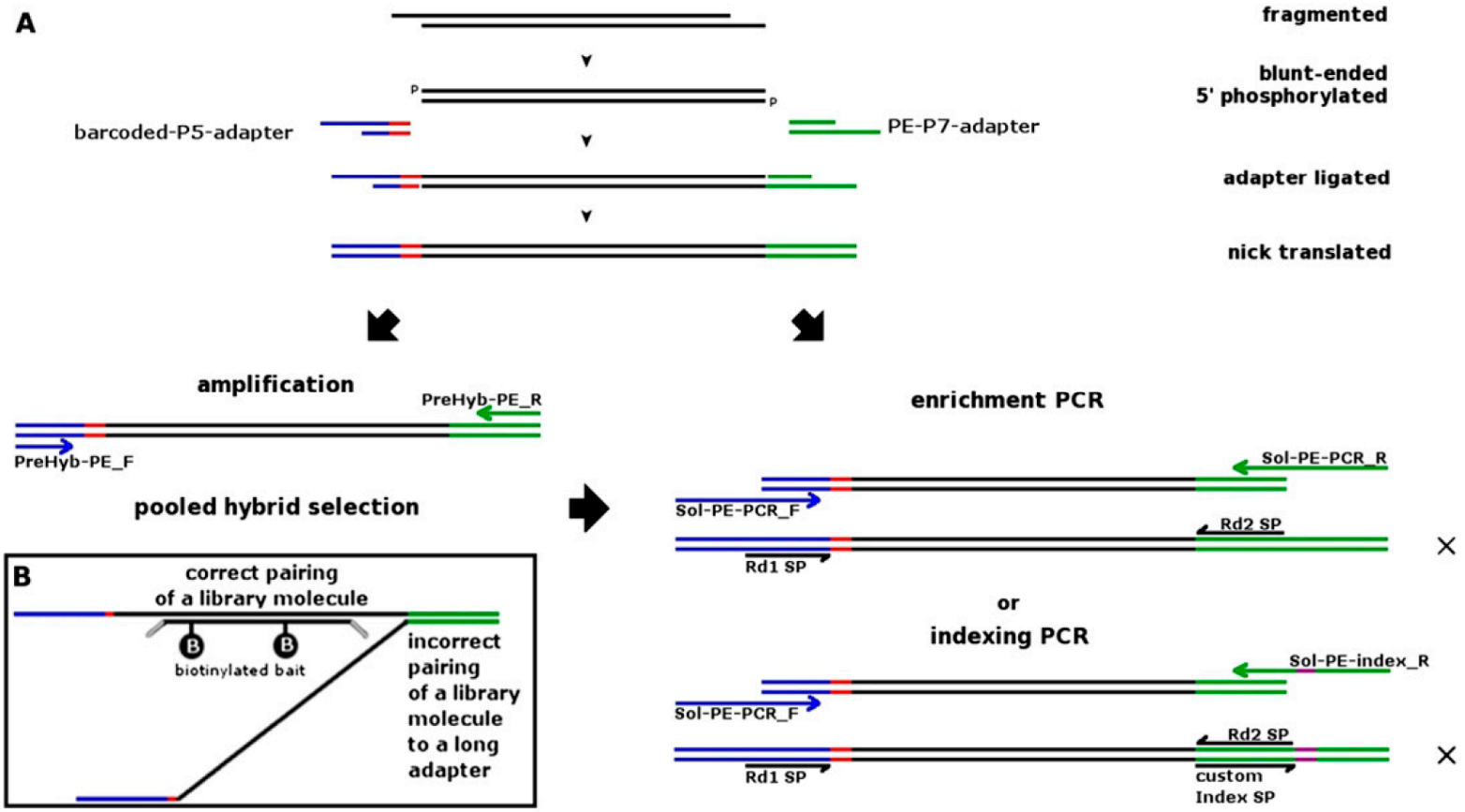
Daisy-chaining

Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture

Nadin Rohland[1] and David Reich

Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA; Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02139, USA

Genome Res. 2012 22: 939-946 originally published online January 20, 2012



Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture

Nadin Rohland[1] and David Reich

Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA; Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02139, USA

Genome Res. 2012 22: 939-946 originally published online January 20, 2012

Pros

- Uses short (33&34bp) adapters with internal barcodes. Barcode is directly ligated to the blunted 5' DNA fragment and become a part of the internal sequence. No need to add barcode-specific blocking oligos in hybridization because adapter sequences are the same for all individual libraries.

- Shorter adapters are less prone to daisy chaining in capture experiments (23% vs 74% capture efficiency using long vs short adapters!).

Cons

- The use of internal barcode will reduce the "effective length" of each read

METHOD Open Access

A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries

Sheila Fisher[1], Andrew Barry[1], Justin Abreu[1], Brian Minie[1], Jillian Nolan[1], Toni M Delorey[1], Geneva Young[1], Timothy J Fennell[1], Alexander Allen[1], Lauren Ambrogio[1], Aaron M Berlin[2], Brendan Blumenstiel[3], Kristian Cibulskis[3], Dennis Friedrich[1], Ryan Johnson[1], Frank Juhn[4], Brian Reilly[1], Ramy Shammas[1], John Stalker[1], Sean M Sykes[2], Jon Thompson[1], John Walsh[1], Andrew Zimmer[1], Zac Zwirko[1,4], Stacey Gabriel[2], Robert Nicol[1], Chad Nusbaum[2*]

Fisher *et al. Genome Biology* 2011, **12**:R1

Pros

- With-beads cleanup during the library prep, which minimizes the loss of DNA samples at each step (80-90% recovery rate compared to 50-60% using standard protocol).
- It saves so much time in the cleanup steps when you work with many samples simultaneously!
- Ideal for low quantity input DNA (e.g. historic DNA).

A General Workflow for *de novo* Exon Capture

1. Target selection and probe design

2. Library preparation and multiplexing

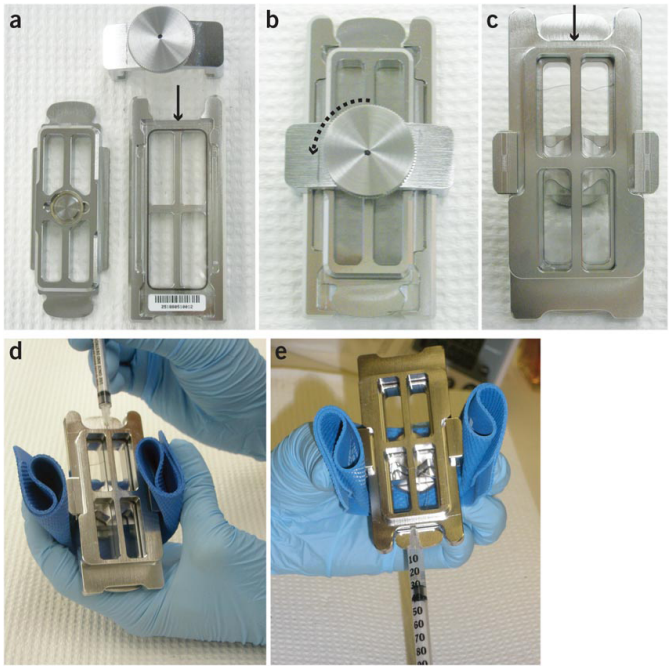
3. Exon capture experiments

4. Post-capture quality control

Protocol for Microarray-based Exon 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.



Tips

- Follow the protocol from Step 29 – Step 61
- Especially pay attention to the critical steps

▲ CRITICAL STEP

- Read table 2 (trouble shooting)
- **Practice on chamber assembly and syringe extraction several times before real experiments**

Protocol for In-solution-based Exon Capture (commercial)

NimbleGen

SeqCap EZ Library SR
User's Guide

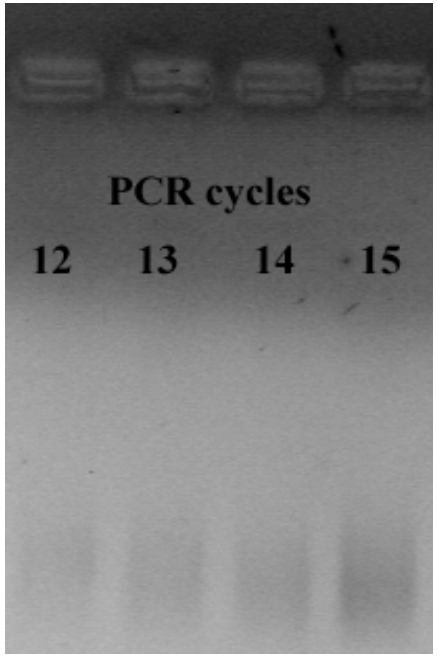
Version 4.2

Tips

- Testing PCR tubes for evaporation
- Follow the protocol from Chapter 5 "Hybridizing the Sample and SeqCap EZ Libraries" to Chapter 7 "Amplifying Captured Multiplex DNA Sample"
- Read carefully each step of the protocol before doing the captures. During the experiment you have to proceed fairly quickly at some steps
- Post-capture amplification: different PCR primers are used for custom library preparation (e.g. IS5 & IS6 for Meyer & Kircher protocol).

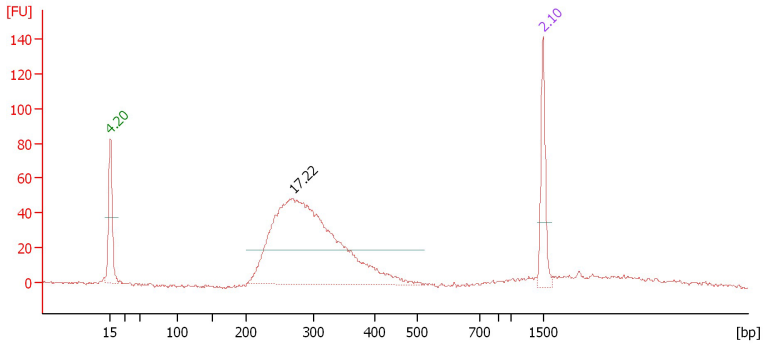
Post Capture Enrichment PCR – Avoid Over Amplification!

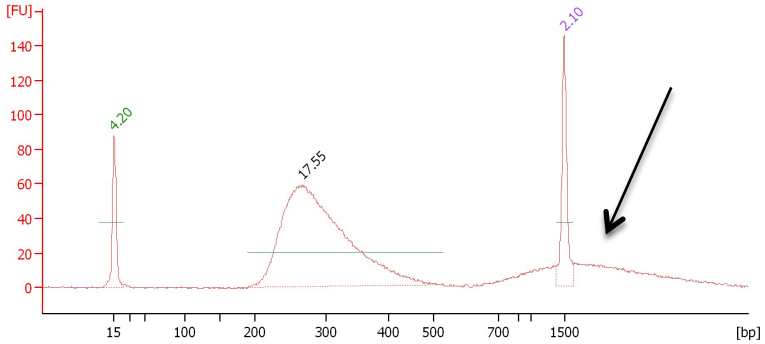
- In post capture enrichment PCR, there is a high probability of barcode swapping especially after PCR reaches saturation – short adapters can act as primers that may anneal to adapters containing different barcodes. Solution: amplify as few cycles as possible and never let your PCR reach plateau.
-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





A General Workflow for *de novo* Exon Capture

1. Target selection and probe design

2. Library preparation and multiplexing

3. Exon capture experiments

4. Post-capture quality control

Measuring Enrichment Efficiency using qPCR

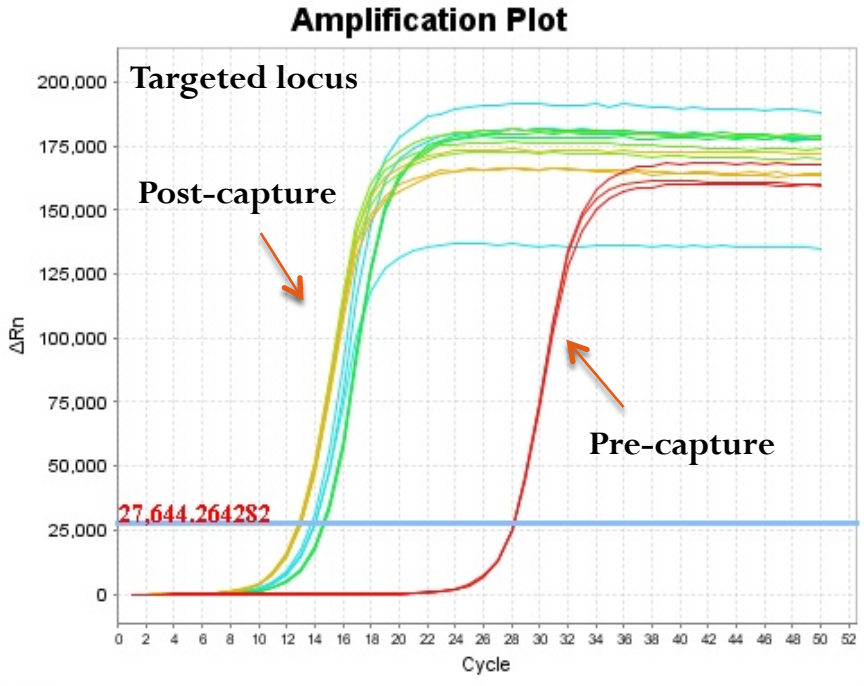
qPCR assays is used to estimate relative fold enrichment by measuring the relative abundance of target loci (positive controls) and non-target loci (negative controls) in pre-capture sample library and post-capture libraries. These assays are an inexpensive way to determine whether the capture was successful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

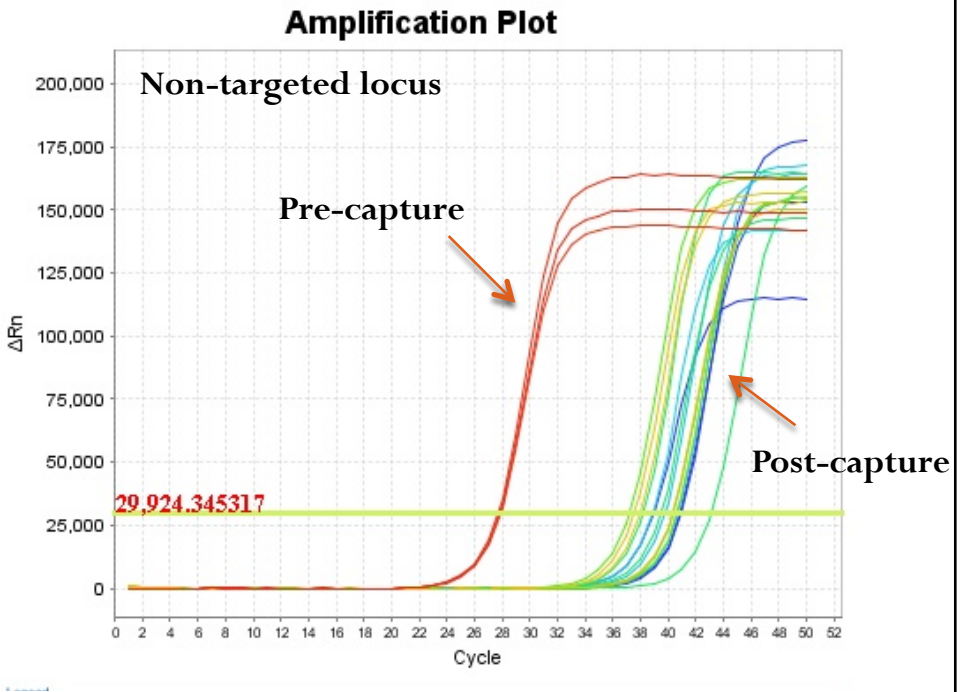
qPCR Primer Design:

- Include 5 positive control and 2-3 negative control loci in your design
- 100-150bp segment of positive controls and negative controls
- Target the central regions
- Tm = 60°C
- GC% 40-60
- Make sure to test your primers before exon capture experiments

Measuring Enrichment Efficiency using qPCR

qPCR assays is used to estimate relative fold enrichment by measuring the relative abundance of target loci (positive controls) and non-target loci (negative controls) 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.

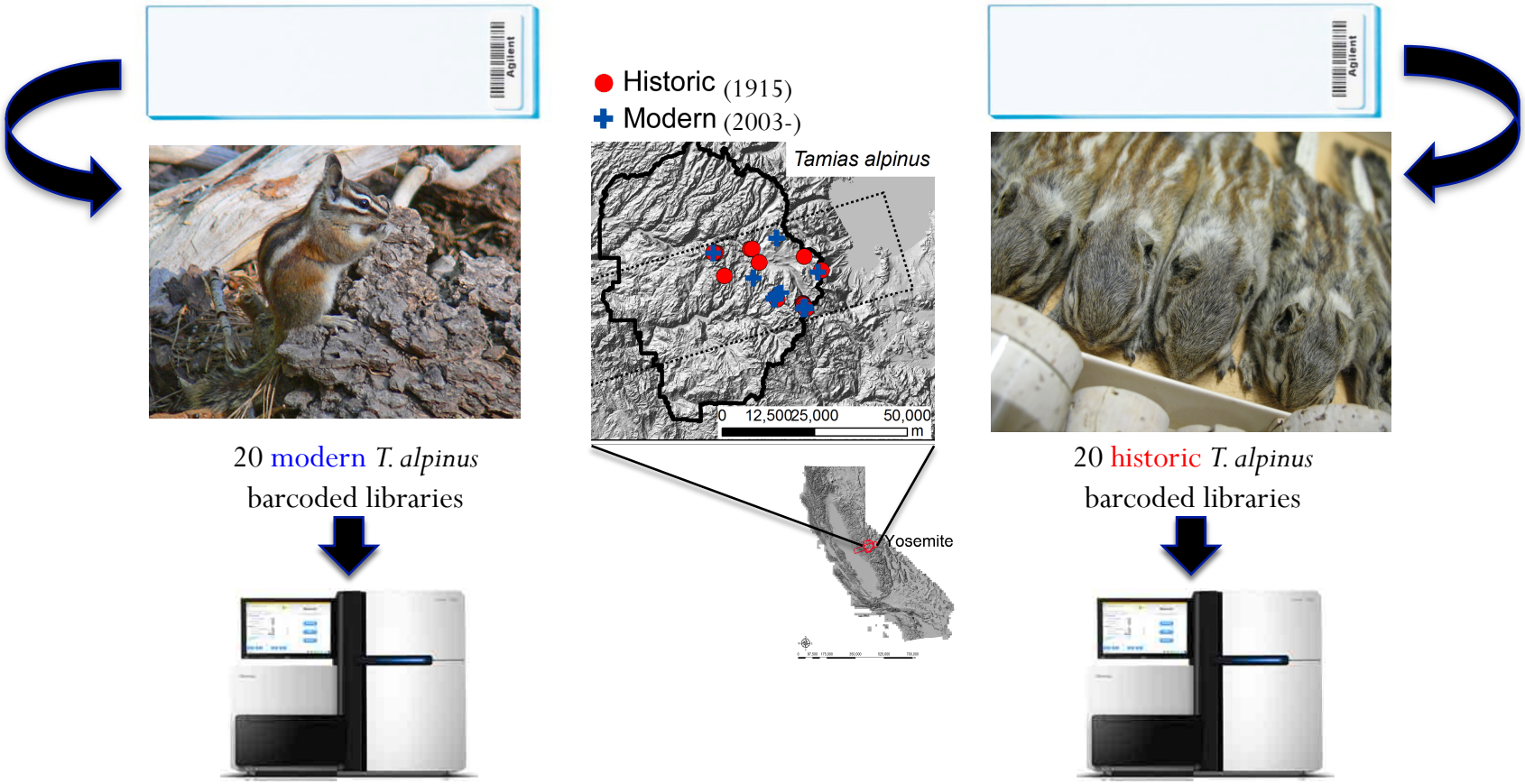




Targeted loci:
Microarray based exon capture: 4-6 cycles' difference
In-solution based exon capture: >8 cycles' difference

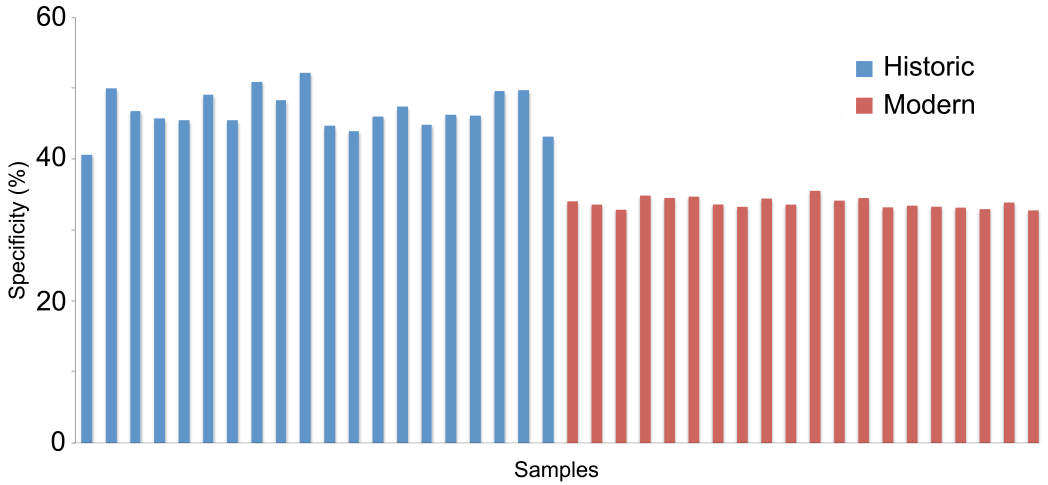
Performance of Transcriptome-based Exon Capture in a Case Study (Array-based)

- **Target exons:** 11,975 exons (>200bp) from ~6000 protein coding genes (4Mb)
- **Quality Control:** 1) *Tamias* mt ; 2) Ground squirrel SRY gene (Y-linked)
- **qPCR:** 7 nuclear control nuclear genes

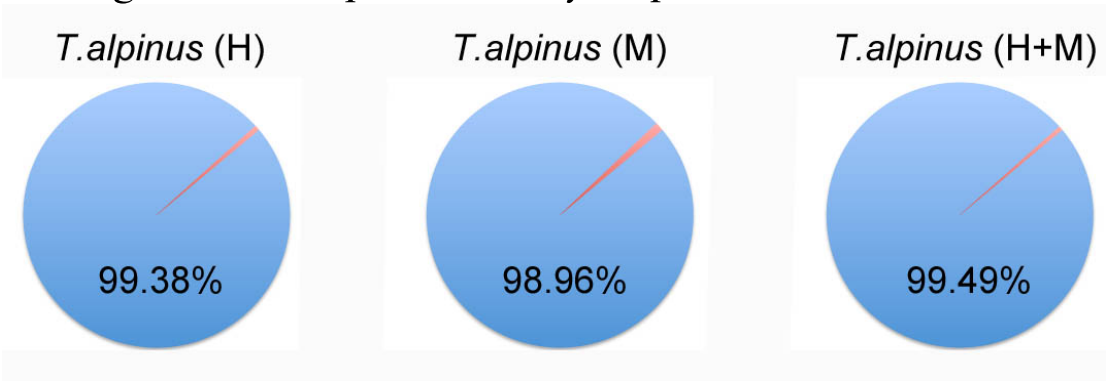


Illumina HiSeq2000, 100PE, 1 lane for each pooled library

Specificity - % cleaned reads mapped to the intended exons



Sensitivity - % target exons represented by sequence reads



	mean(bp)	median(bp)	N50(bp)	length(Mb)
Target exons	**332**	**245**	**308**	**3.99**
In-target assemblies	**715**	**595**	**722**	**7.58**

In-target assemblies = target exons + flanking sequences.

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

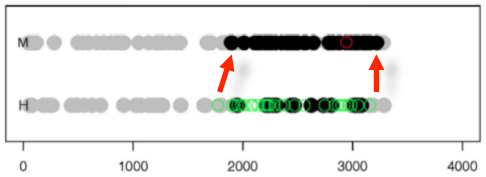
Total target size: 9.32 Mb

- ~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

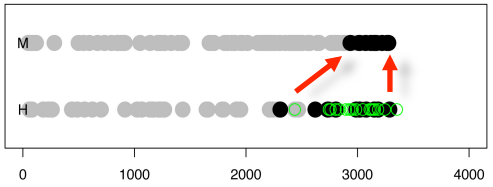


Modern: N=48

Historic: N=56

Retracting at Yosemite

T. alpinus

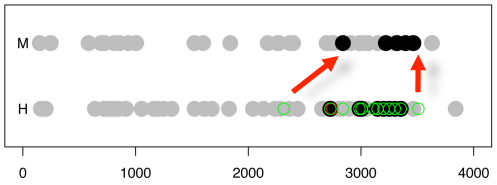


Modern: N=48

Historic: N=55

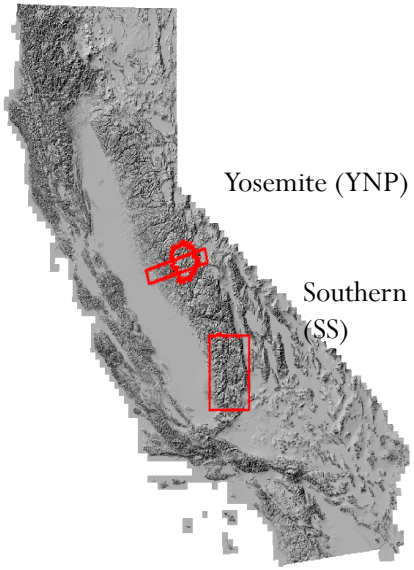
Retracting at southern Sierra

T. alpinus



Modern: N=41

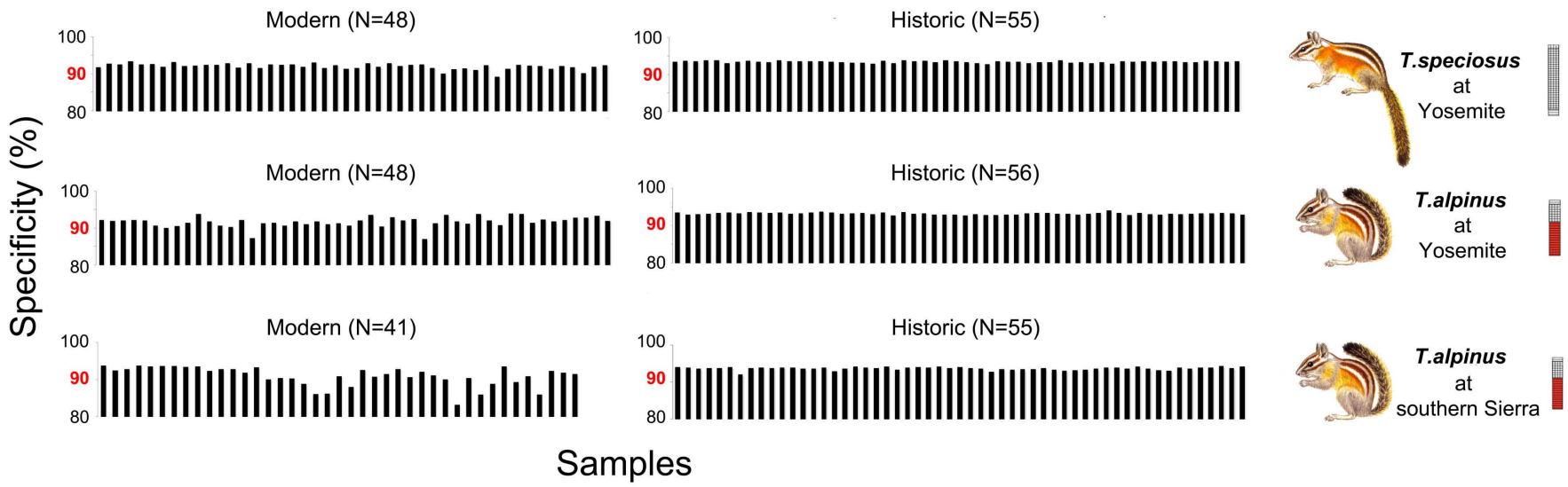
Historic: N=55



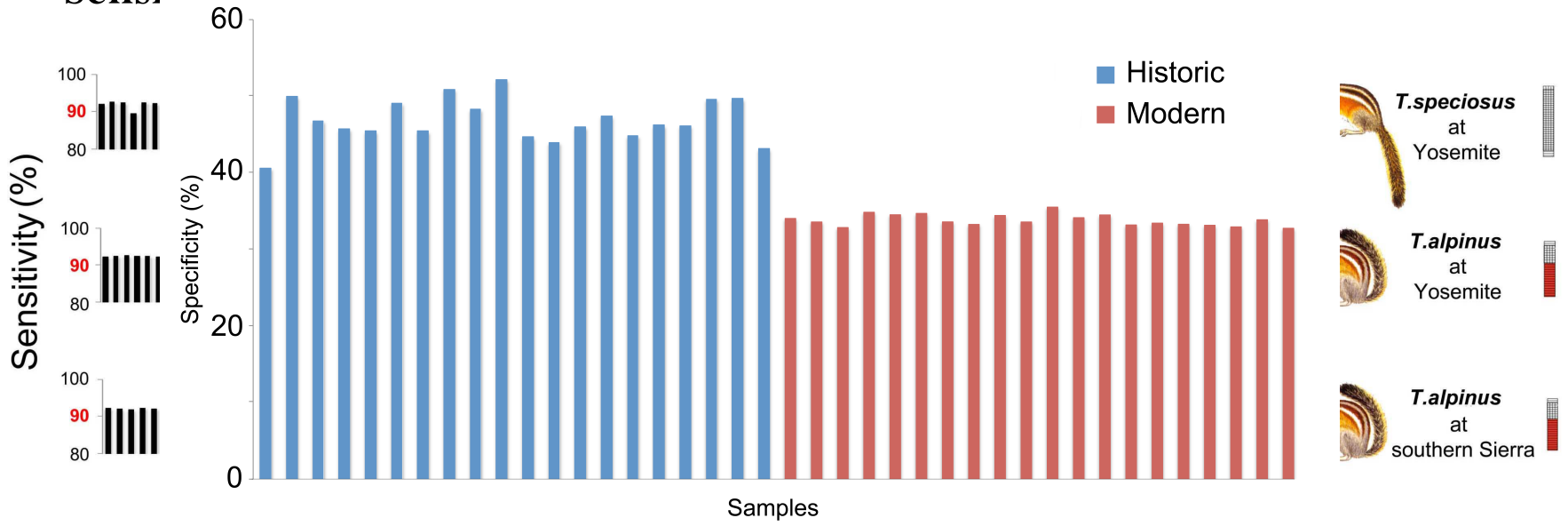
NimbleGen in-solution capture & sequencing:

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

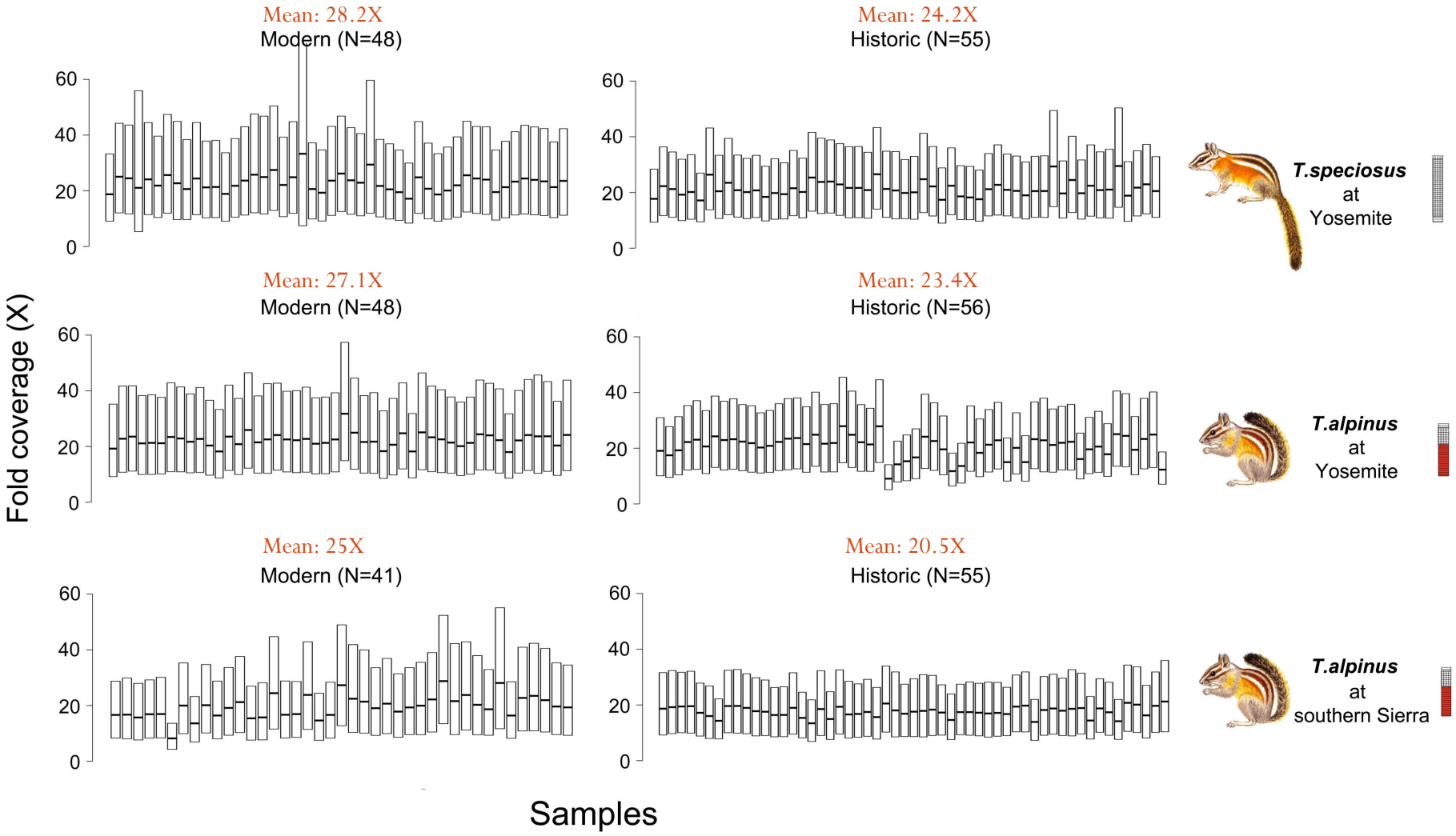
Specificity - % cleaned reads mapped to the intended exons



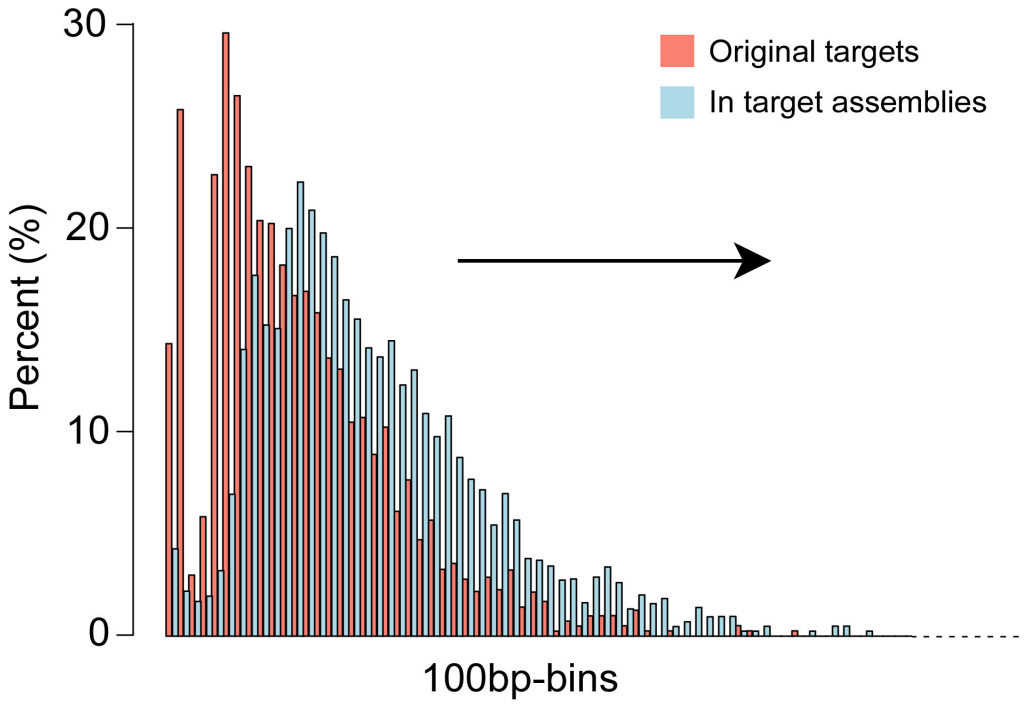
Sensi



Average Sequence Depth



In-target Assemblies vs. Target Exons



	Total length(Mbp)	Mean(bp)	Median(bp)	N50(bp)
Original Targets	9.32	512	465	720
In-target assemblies	20.8	986	880	1135

Bioinformatics Pipelines for Transcriptome-Based Exon Capture

- Data cleanup- trimming for quality, removing adapters, merging overlapping reads, removing duplicates and reads sourced from contamination
- *De novo* assembly of reads across multiple individuals using multiple k-mers
- Merge contigs across your multiple k-mer assemblies to reduce redundancy
- Identify targeted loci in *de novo* assembly by doing a BLAST search
- Align individual reads to the targeted loci assembly
- Reconstruct orthologous loci (phylogenomics) and call variants (population genomics)
- Data filtering
- Population genomic and phylogenetic analyses

Pipeline repositories:

- *de novo* transcritome assembly, annotation, and marker selection
https://github.com/MVZSEQ/denovoTranscriptomeMarkerDevelopment
- *de novo* exon captue data analyses for population genomics
https://github.com/MVZSEQ/denovoTargetCapture

Pipelines under development

- *de novo* exon captue data analyses for Phylogenomics
- UCE data analyses for Phylogenomics
- PopGenTools: Population genomics analyses using ANGSD & ngsTools
- RADTools