

Quality Assessment of Hybrid Selection Experiments

Kristian Cibulskis
Andrew Kernytsky

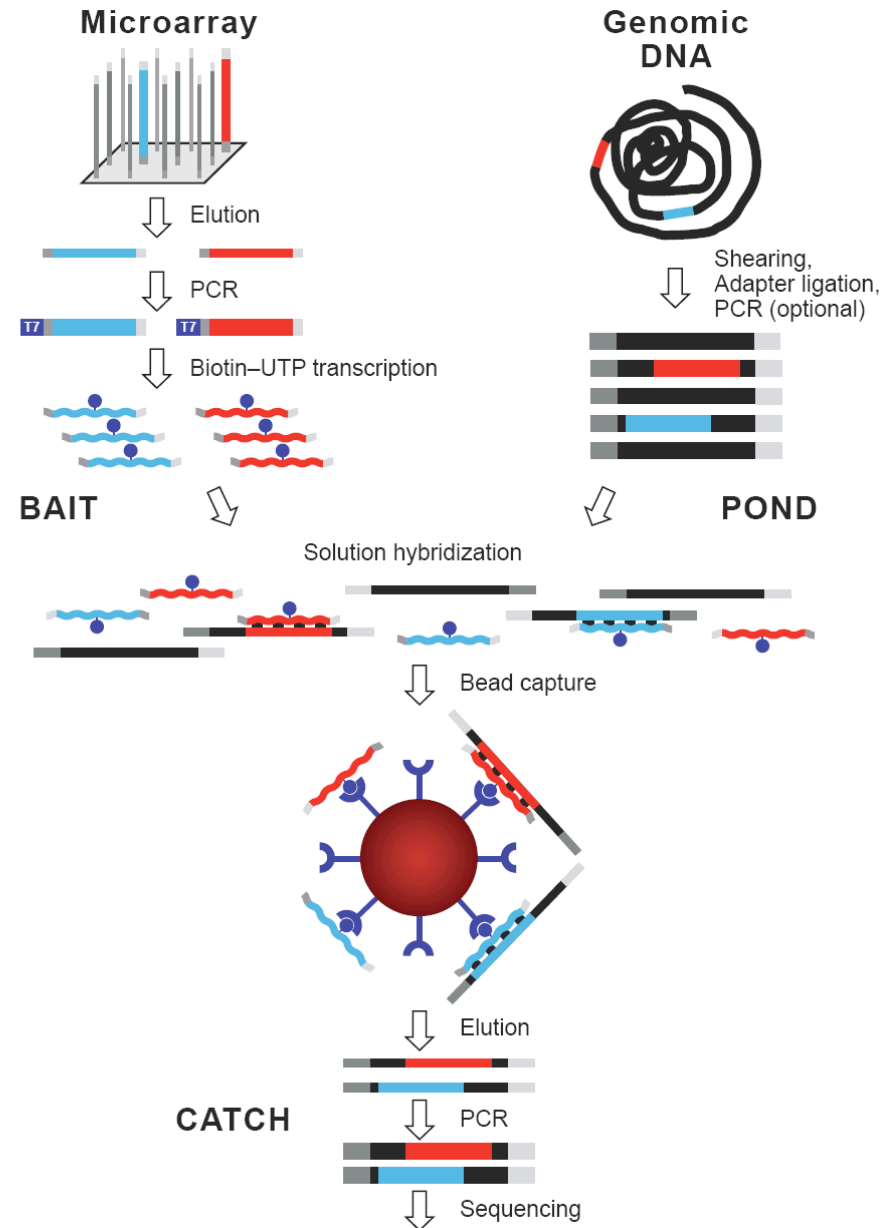
Genome Sequencing and Analysis
Broad Institute of Harvard and MIT
02/04/10

Broad Capture Process

Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing

Andreas Gnirke¹, Alexandre Melnikov¹, Jared Maguire¹, Peter Rogov¹, Emily M LeProust², William Brockman^{1,5}, Timothy Fennell¹, Georgia Giannoukos¹, Sheila Fisher¹, Carsten Russ¹, St David B Jaffe¹, Eric S Lander^{1,3,4} & Chad Nusbaum¹

VOLUME 27 NUMBER 2 FEBRUARY 2009 **NATURE BIOTECHNOLOGY**

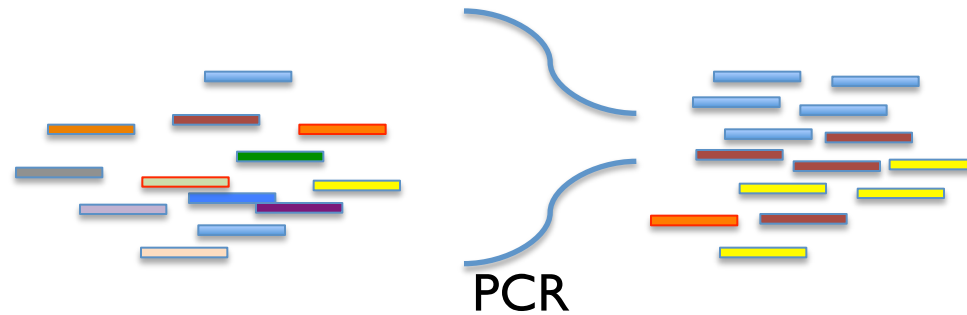


Molecular Duplicates

ISSUE: Molecular Duplicates

Problem

- PCR-free protocol yields small amounts of DNA
- Bottlenecks decrease library complexity



```
AAAGCCTGGAGGTACAGCTTCAGGAGCTGCTTGGACTGACGCACCTCCTCTCTTTGGTCAAGCACCTGTTGGAGTG
AAAGCCTGGAGGTACAGCTGCAGGAGCTGCTTGGACTGACGCACCTCCTCTCTTTGGTCAAGCACCTGTTGGAGTG
AAAGCCTGGAGGTACAGCTGCAGGAGCTGCTTGGACTGACGCACCTCCTCTCTTTGGTCAAGCACCTGTTGGAGTG
CCTGGAGGTACAGCTGCAGGAGCTGCTTGGACTGACGCACCTCCTCTCTTTGGTCAAGCACCTGTTGGAGTGTGTG
GAGCCGCTTGGACTGACGCACCTCCTCTCTTTGGTCAAGCACCTGTTGGAGTGTGTGCTGCAGCCGCTGGACGGTG
GAGCTGCTTGGACTGACGCACCTCCTCTCTTTGGTCAAGCACCTGTTGGAGTGTGTGCTGCAGCCGCTGGACGGTG
CGCACTTCCTCTCTTTGGTCAAGCACCTGTTGGAGTGTGTGCTGCAGCCGCTGGACGGTGGCACAACCTCTGACGTA
CGCACTTCCTCTCTTTGGTCAAGCACCTGTTGGAGTGTGTGCTGCAGCCGCTGGACGGTGGCACAACCTCTGACGTA
CGCACTTCCTCTCTTTGGTCAAGCACCTGTTGGAGTGTGTGCTGCAGCCGCTGGACGGTGGCACAACCTCTGACGTA
CGCACTTCCTCTCTTTGGTCAAGCACCTGTTGGAGTGTGTGCTGCAGCCGCTGGACGGTGGCACAACCTCTGACGTA
CCTCTCTTTGGTCAAGCACCTGTTGGAGTGTGTGCTGCAGCCGCTGGACGGTGGCACAACCTCTGACGTAAGTTCCTG
CCTCTCTTTGGTCAAGCACCTGTTGGAGTGTGTGCTGCAGCCGCTGGACGGTGGCACAACCTCTGACGTAAGTTCCTG
CTCTTTGGTCAAGCACCTGTTGGAGTGTGTGCTGCAGCCGCTGGACGGTGGCACAACCTCTGACGTAAGTTCCTGAGC
```

Running MarkDuplicates

- *MarkDuplicates*
 - Part of the Picard toolset and pipeline
 - By default mark, not remove, duplicate molecules
 - Automatically run on every Hybrid Selection library
 - Results located in the per-library directory below your aggregated BAM file:
 - E.g. /seq/picard_aggregation/C278/18325/v1/Solexa-16035/
Solexa-16035.duplicate_metrics

```
java -Xmx2048m -jar /seq/software/picard/current/bin/MarkDuplicates.jar  
INPUT=<Solexa-16035.bam>  
OUTPUT=<Solexa-16035.duplicates_marked.bam>  
METRICS=<Solexa-16035.duplicate_metrics>
```

Output Metrics text file with metrics
about the duplicate marking process

Output BAM with duplicate
records marked

See <http://picard.sourceforge.net/command-line-overview.shtml#MarkDuplicates> for more information

Interpreting MarkDuplicates

- **Percent Duplication**

- *Seems* easy to understand and communicate
- **Increases with sequencing depth**

- **Estimated Library Size**

- Estimate of number of unique molecules in the library
- Goal changes with target size
 - Exome - ~150M
 - 6000 Genes – ~50M

- ***No increase with sequencing depth***

Metric	Value
LIBRARY	Solexa-16035
UNPAIRED_READS_EXAMINED	948,391
READ_PAIRS_EXAMINED	46,752,523
UNMAPPED_READS	17,451,483
UNPAIRED_READ_DUPLICATES	621,011
READ_PAIR_DUPLICATES	6,929,586
PERCENT_DUPLICATION	15.3%
ESTIMATED_LIBRARY_SIZE	141,711,971

Evaluating Your Selection Event

Running Hybrid Selection Metrics

- *CalculateHSMetrics*
 - Part of the Picard toolset and pipeline
 - Automatically run on every HS lane, library and aggregation
 - Results located next to your BAM file:
 - E.g. /seq/picard_aggregation/C278/18325/v1/18325.duplicate_metrics

```
java -Xmx2048m -jar /seq/software/picard/current/bin/CalculateHsMetrics.jar  
  BAIT_INTERVALS=<whole_exome_agilent_designed_120.baits.interval_list>  
  TARGET_INTERVALS=<whole_exome_agilent_designed_120.targets.interval_list>  
  INPUT=18325.bam>  
  OUTPUT=<18325.hybrid_selection_metrics>
```

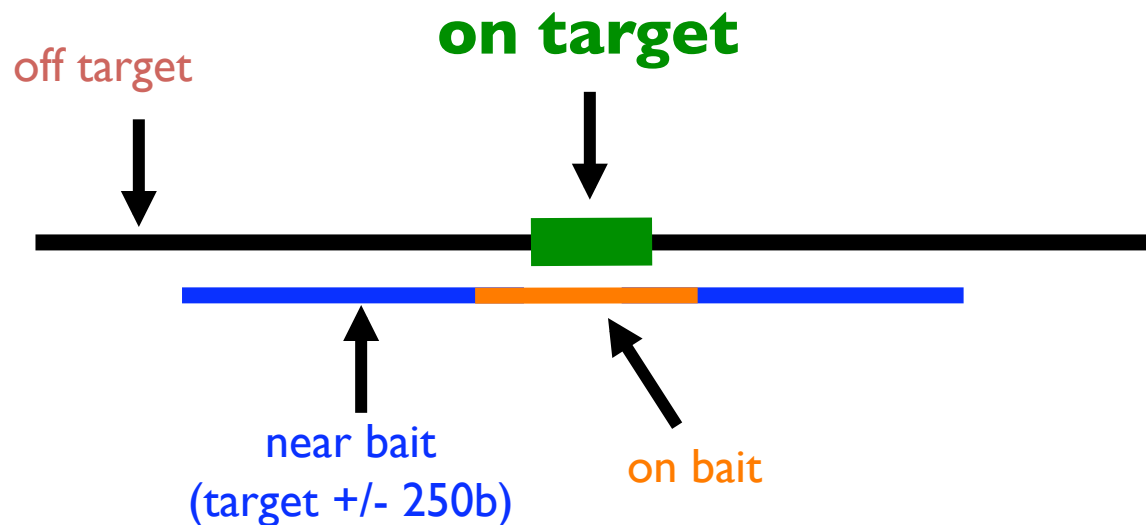
Output text file with metrics about this
Hybrid Selection event

Target and Bait definitions for the
selection to evaluate

Interpreting HS Metrics

- **Percent Selected Bases**

- Percentage of the PF aligned bases were on or near the target
- Varies by target size
 - Exome: good selections > 80%



Sample: 18325 (Microcephaly Exome)	
Project: C278	
Sample Status: Complete	
Bait Territory	37,640,396
Target Territory	28,646,006
Bait Design Efficiency	76.10%
Total Reads	111,904,920
PF Reads	96,322,338
% Unique Reads	81,842,155
% PF Reads	86.08%
PF Unique Bases Aligned	5,310,101,316
% Selected Bases	80.38%
Mean Bait Coverage	95.71
Mean Target Coverage	104.50
% Zero Coverage Targets	2.68%
Fold 80 Base Penalty	4.35
Target Bases >= 2x	94.68%
Target Bases >= 10x	88.02%
Target Bases >= 20x	80.95%
Target Bases >= 30x	74.60%
HS Penalty 10x	8.63
HS Penalty 20x	8.97
HS Penalty 30x	9.76

See http://www.broadinstitute.org/~prodinfo/picard_metric_definitions.html for more information

Interpreting HS Metrics

- **% Zero Coverage Targets**
 - Percentage of targets that don't have at least one base $\geq 2X$
 - Often are targets that are difficult to *sequence*, as in whole genomes these are missing also (repetitive, high gc)
 - For exomes, typically $< 3\%$

Sample: 18325 (Microcephaly Exome) Project: C278 Sample Status: Complete	
Bait Territory	37,640,396
Target Territory	28,646,006
Bait Design Efficiency	76.10%
Total Reads	111,904,920
PF Reads	96,322,338
% Unique Reads	81,842,155
% PF Reads	86.08%
PF Unique Bases Aligned	5,310,101,316
% Selected Bases	80.38%
Mean Bait Coverage	95.71
Mean Target Coverage	104.50
% Zero Coverage Targets	2.68%
Fold 80 Base Penalty	4.35
Target Bases $\geq 2x$	94.68%
Target Bases $\geq 10x$	88.02%
Target Bases $\geq 20x$	80.95%
Target Bases $\geq 30x$	74.60%
HS Penalty 10x	8.63
HS Penalty 20x	8.97
HS Penalty 30x	9.76

See http://www.broadinstitute.org/~prodinfo/picard_metric_definitions.html for more information

Interpreting HS Metrics

- **Mean Bait / Target Coverage**
- **% of Target Bases > 2x, 10x, 20x, 30x**
 - Measure of project completeness
 - Typically 80% of target over 20x
- Difference between these two illustrates non-uniformity in coverage...

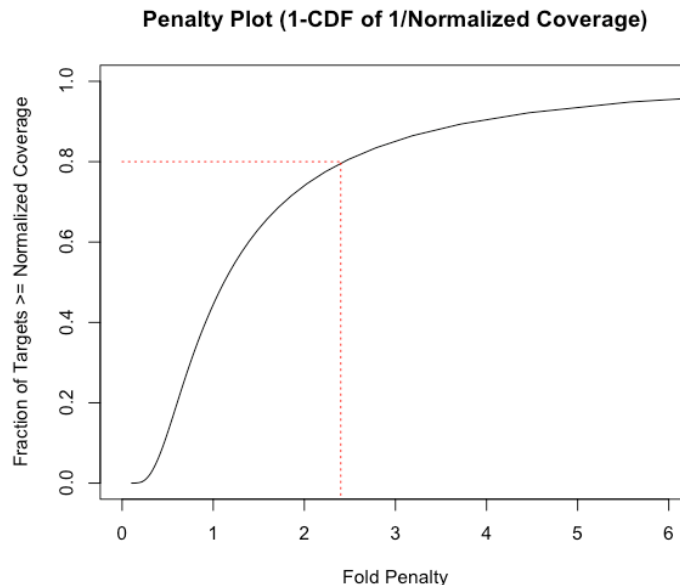
Sample: 18325 (Microcephaly Exome)	
Project: C278	
Sample Status: Complete	
Bait Territory	37,640,396
Target Territory	28,646,006
Bait Design Efficiency	76.10%
Total Reads	111,904,920
PF Reads	96,322,338
% Unique Reads	81,842,155
% PF Reads	86.08%
PF Unique Bases Aligned	5,310,101,316
% Selected Bases	80.38%
Mean Bait Coverage	95.71
Mean Target Coverage	104.50
% Zero Coverage Targets	2.68%
Fold 80 Base Penalty	4.35
Target Bases >= 2x	94.68%
Target Bases >= 10x	88.02%
Target Bases >= 20x	80.95%
Target Bases >= 30x	74.60%
HS Penalty 10x	8.63
HS Penalty 20x	8.97
HS Penalty 30x	9.76

See http://www.broadinstitute.org/~prodinfo/picard_metric_definitions.html for more information

Interpreting HS Metrics

- **Fold 80 Base Penalty**

- Measure of non-uniformity
- # of units of sequencing to raise 80% of the bases to the average coverage
- Lower is better, less than 5 is good and exomes are often between 3-4



Sample: 18325 (Microcephaly Exome)	
Project: C278	
Sample Status: Complete	
Bait Territory	37,640,396
Target Territory	28,646,006
Bait Design Efficiency	76.10%
Total Reads	111,904,920
PF Reads	96,322,338
% Unique Reads	81,842,155
% PF Reads	86.08%
PF Unique Bases Aligned	5,310,101,316
% Selected Bases	80.38%
Mean Bait Coverage	95.71
Mean Target Coverage	104.50
% Zero Coverage Targets	2.68%
Fold 80 Base Penalty	4.35
Target Bases >= 2x	94.68%
Target Bases >= 10x	88.02%
Target Bases >= 20x	80.95%
Target Bases >= 30x	74.60%
HS Penalty 10x	8.63
HS Penalty 20x	8.97
HS Penalty 30x	9.76

See http://www.broadinstitute.org/~prodinfo/picard_metric_definitions.html for more information

Interpreting HS Metrics

- **HS Penalty 10x / 20x / 30x**
 - Overall penalty to get 80% of bases to 10x
 - Captures all targeting inefficiencies
 - In other words... how much input sequence do I need to get a target base to 10x
 - Example:
 - 30 mb of target
 - 10x desired coverage
 - $30\text{mb} * 10x * 8.63 \rightarrow 2.59 \text{ GB}$
 - *A good value is < 10*
 - ***Sometimes this is empty, meaning it's impossible to sequence this library to the desired depth***

Sample: 18325 (Microcephaly Exome)	
Project: C278	
Sample Status: Complete	
Bait Territory	37,640,396
Target Territory	28,646,006
Bait Design Efficiency	76.10%
Total Reads	111,904,920
PF Reads	96,322,338
% Unique Reads	81,842,155
% PF Reads	86.08%
PF Unique Bases Aligned	5,310,101,316
% Selected Bases	80.38%
Mean Bait Coverage	95.71
Mean Target Coverage	104.50
% Zero Coverage Targets	2.68%
Fold 80 Base Penalty	4.35
Target Bases >= 2x	94.68%
Target Bases >= 10x	88.02%
Target Bases >= 20x	80.95%
Target Bases >= 30x	74.60%
HS Penalty 10x	8.63
HS Penalty 20x	8.97
HS Penalty 30x	9.76

See http://www.broadinstitute.org/~prodinfo/picard_metric_definitions.html for more information

References

- Bait / Target Information
 - /seq/references/HybSelOligos/...
- Picard Tools
 - <http://picard.sourceforge.net>
 - /seq/software/picard/current/bin
- Picard Documentation
 - <http://picard.sourceforge.net/command-line-overview.shtml>
 - [http://www.broadinstitute.org/~prodinfo/picard metric definitions.html](http://www.broadinstitute.org/~prodinfo/picard_metric_definitions.html)