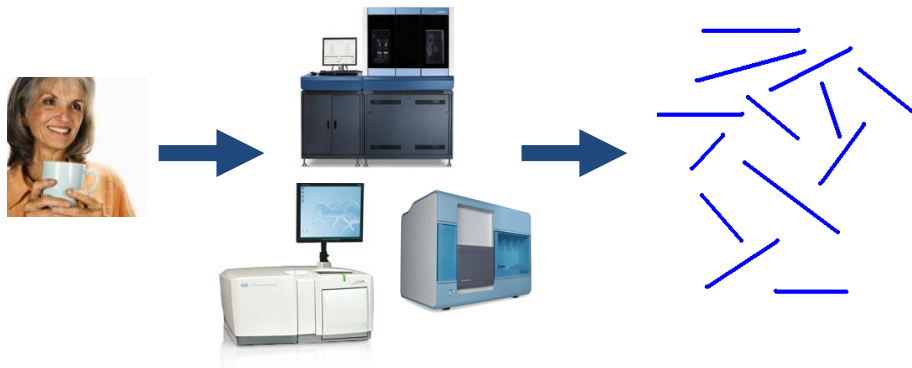


Target Enrichment

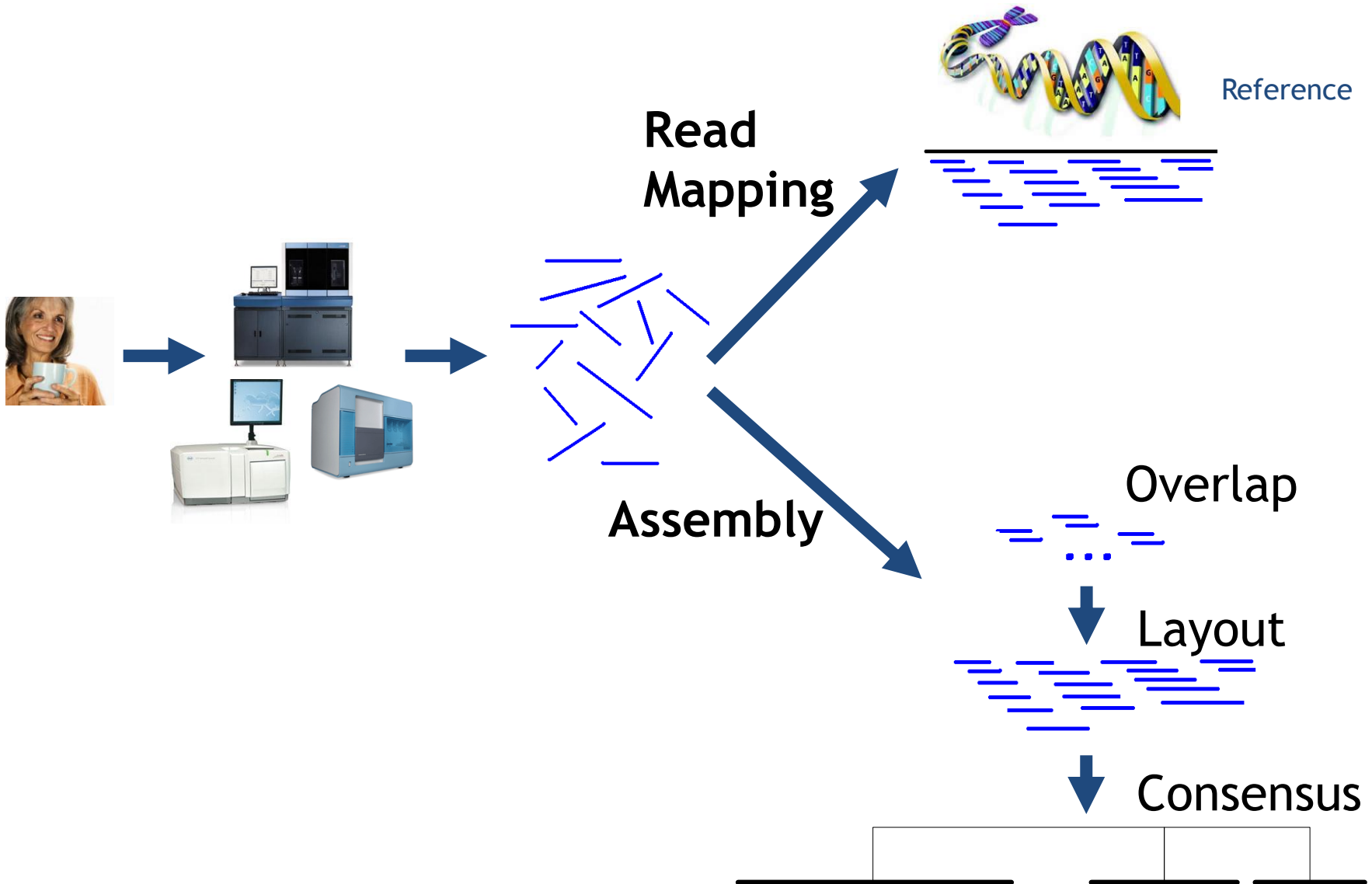
Tobias Rausch

June 2011

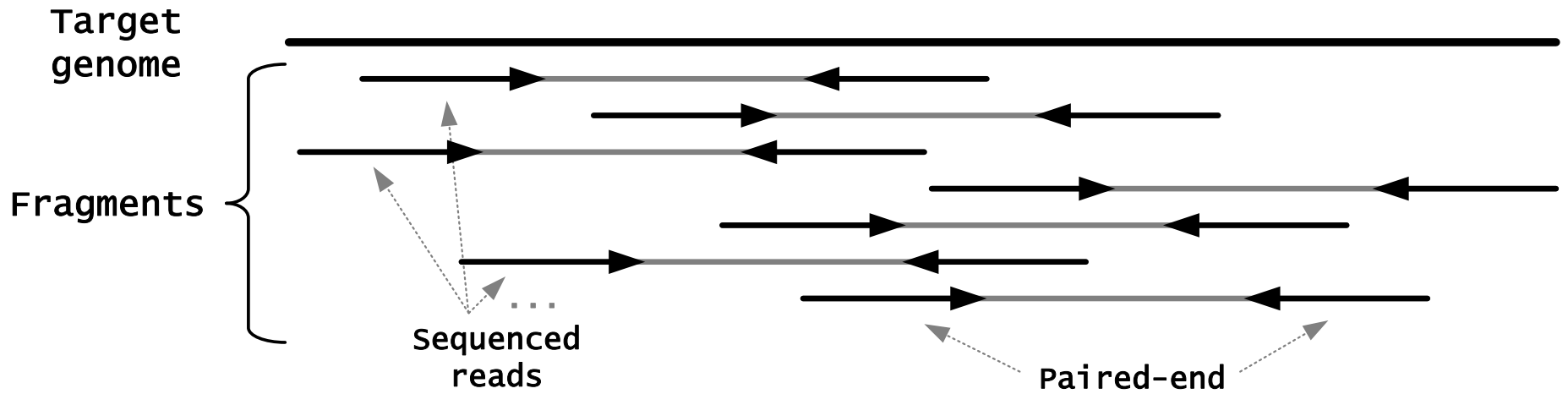
Data Analysis



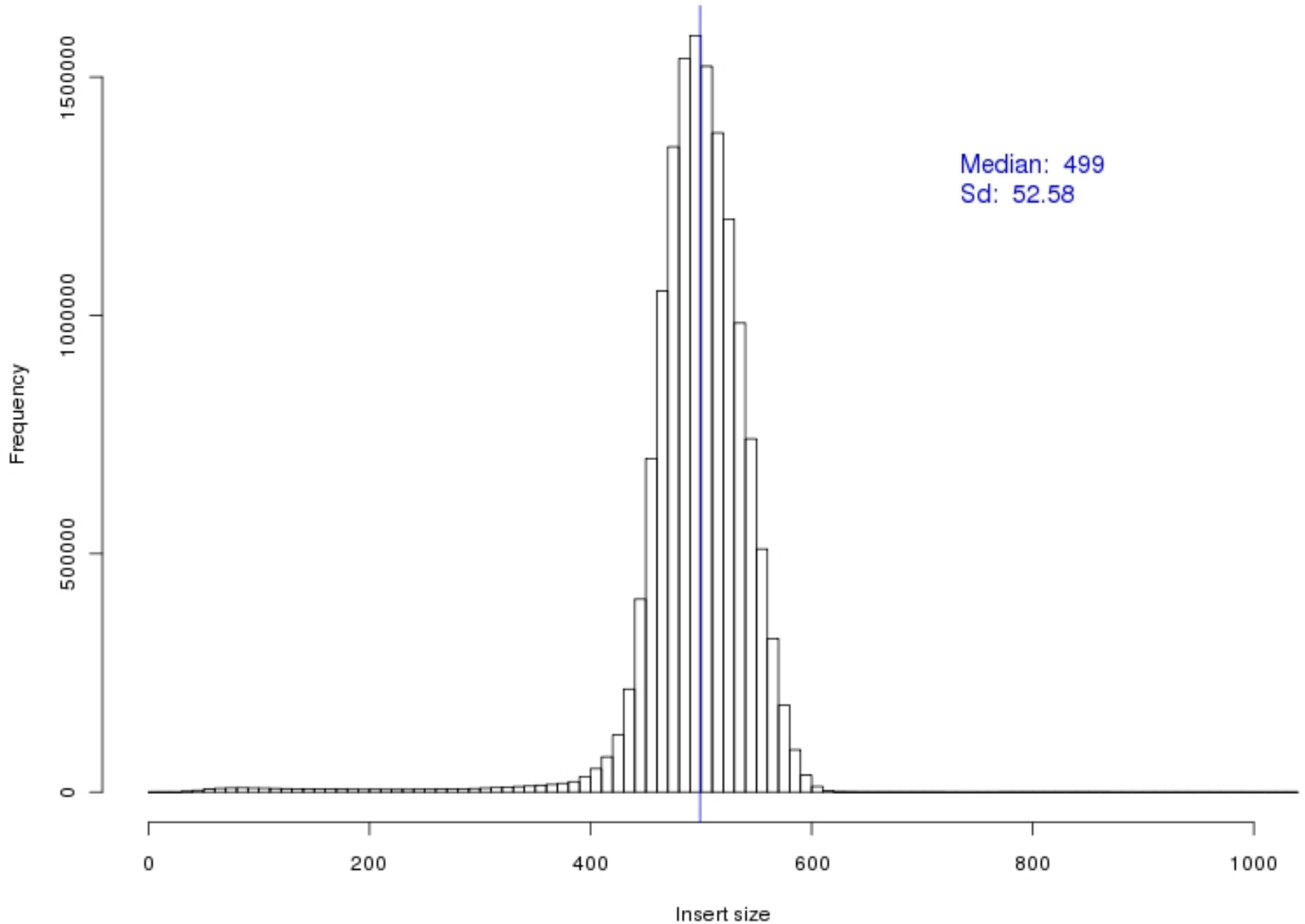
Data Analysis



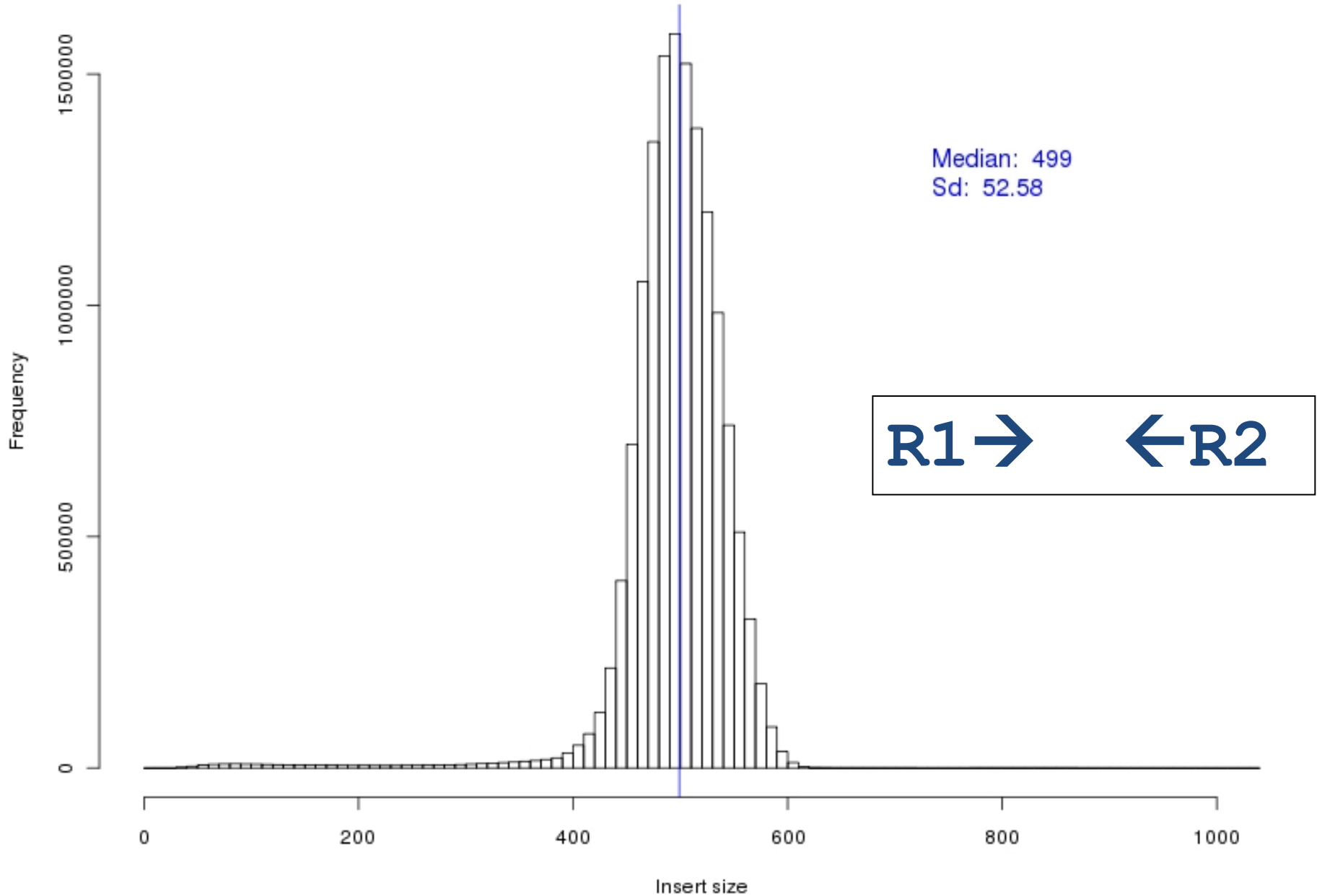
Paired-End Sequencing



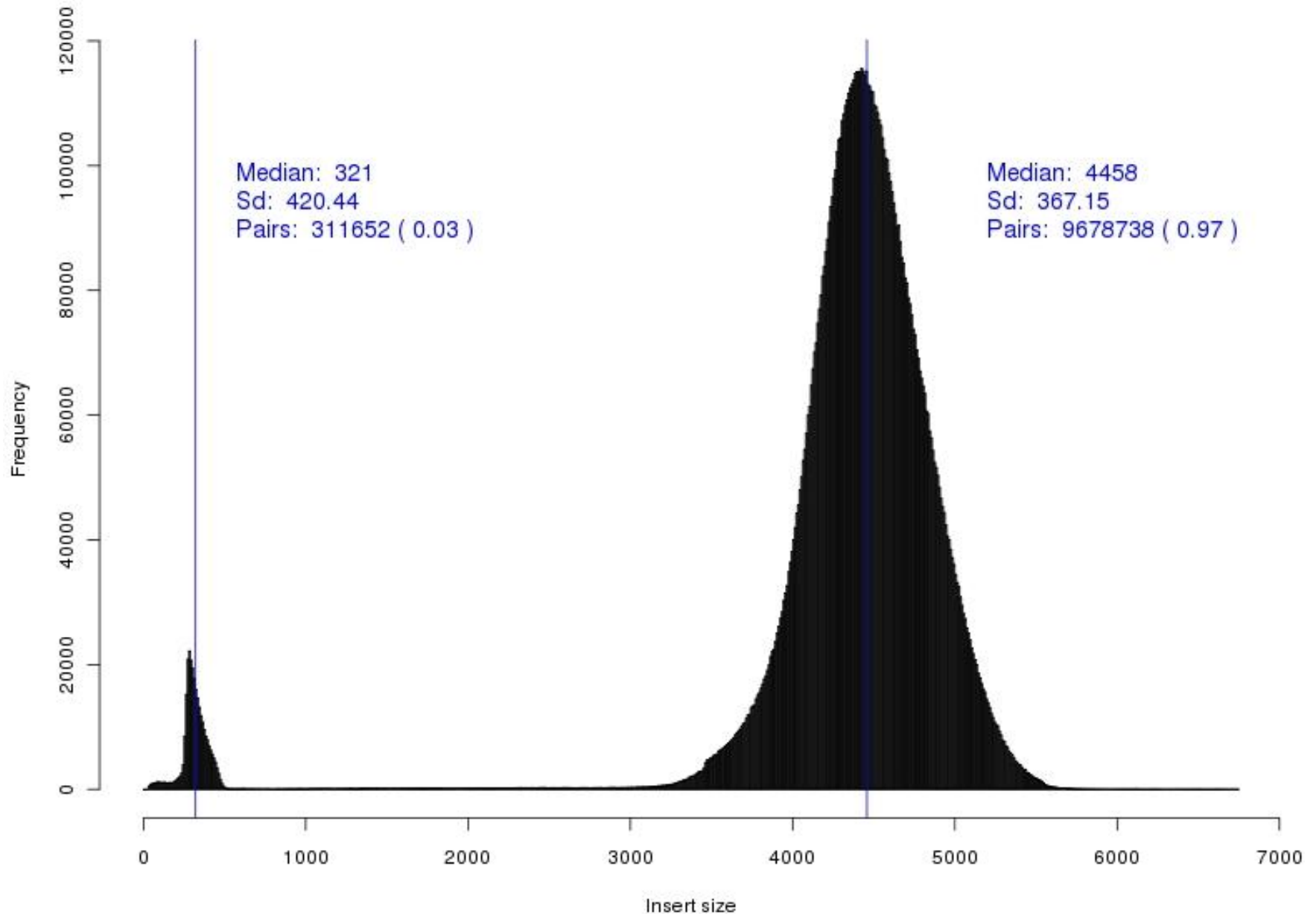
Paired-End Libraries



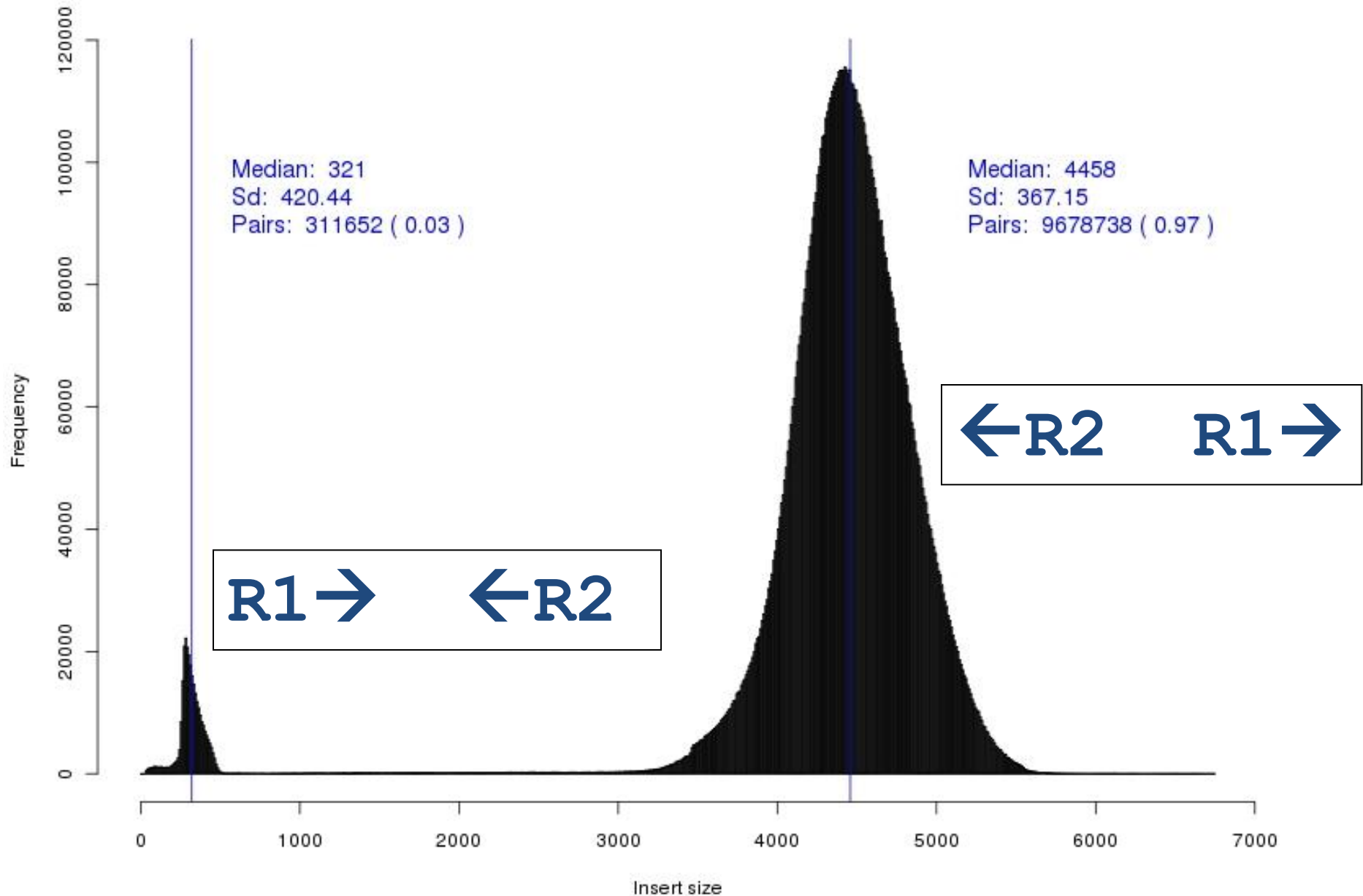
Paired-End Libraries



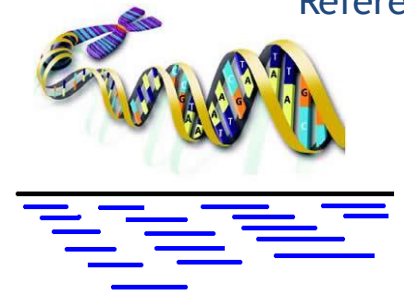
Mate-Pair Libraries



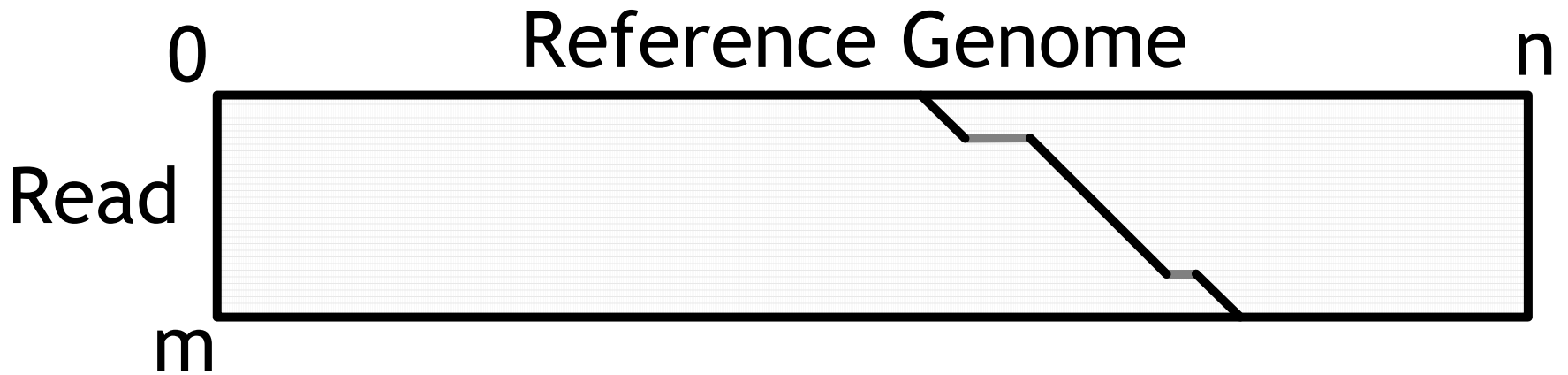
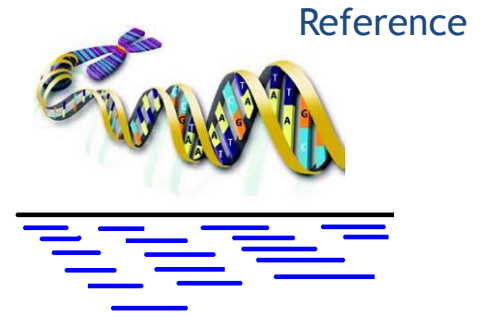
Mate-Pair Libraries



Read Mapping

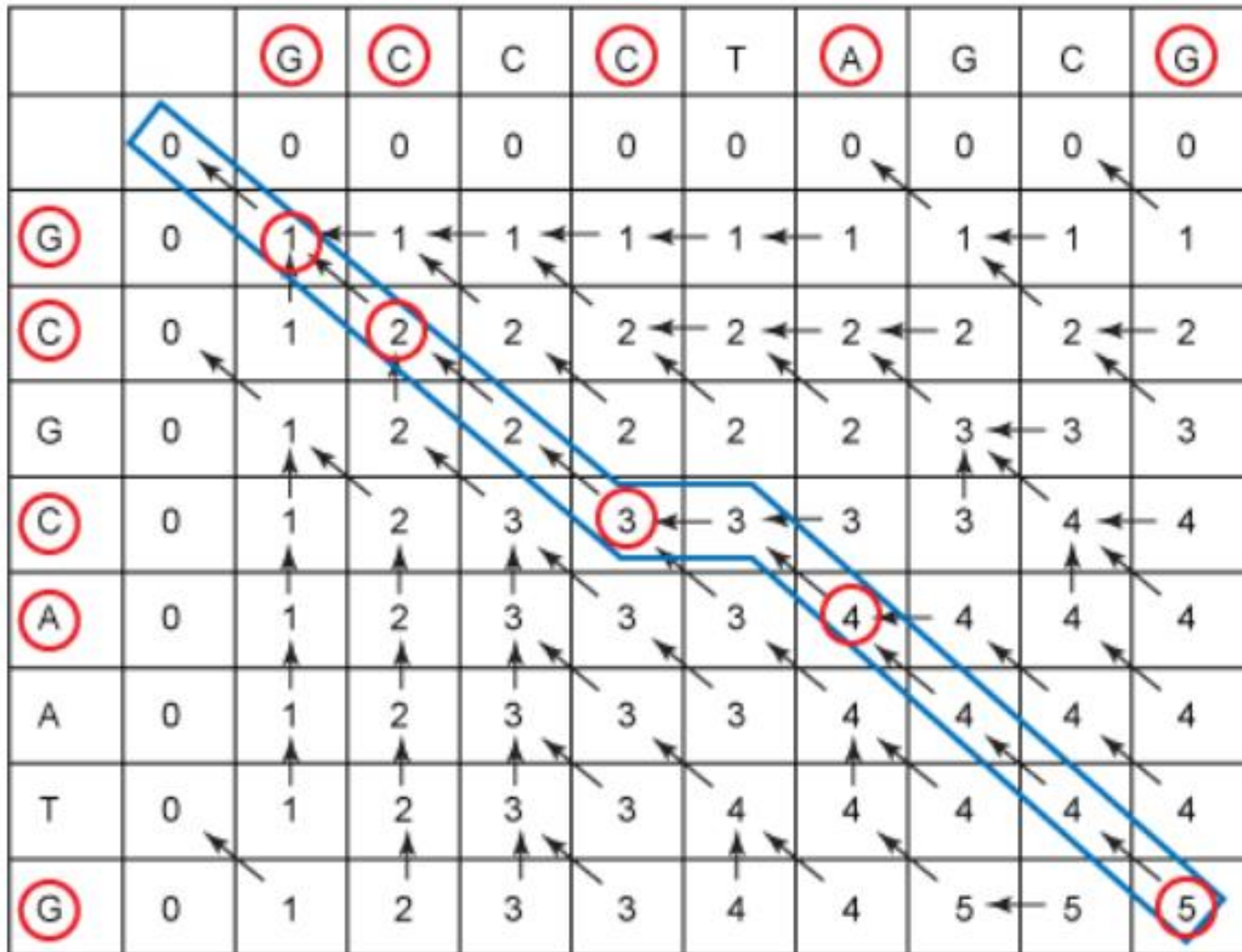
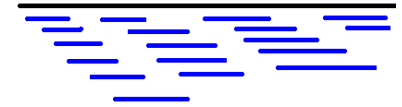


Read Mapping





Read Mapping



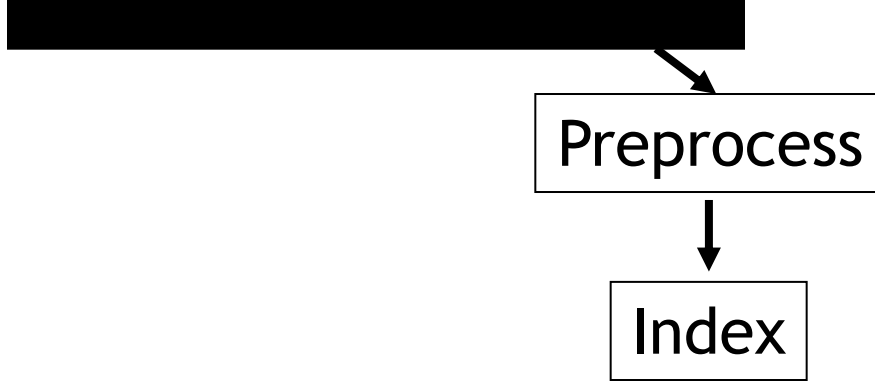
Filtering

Genome

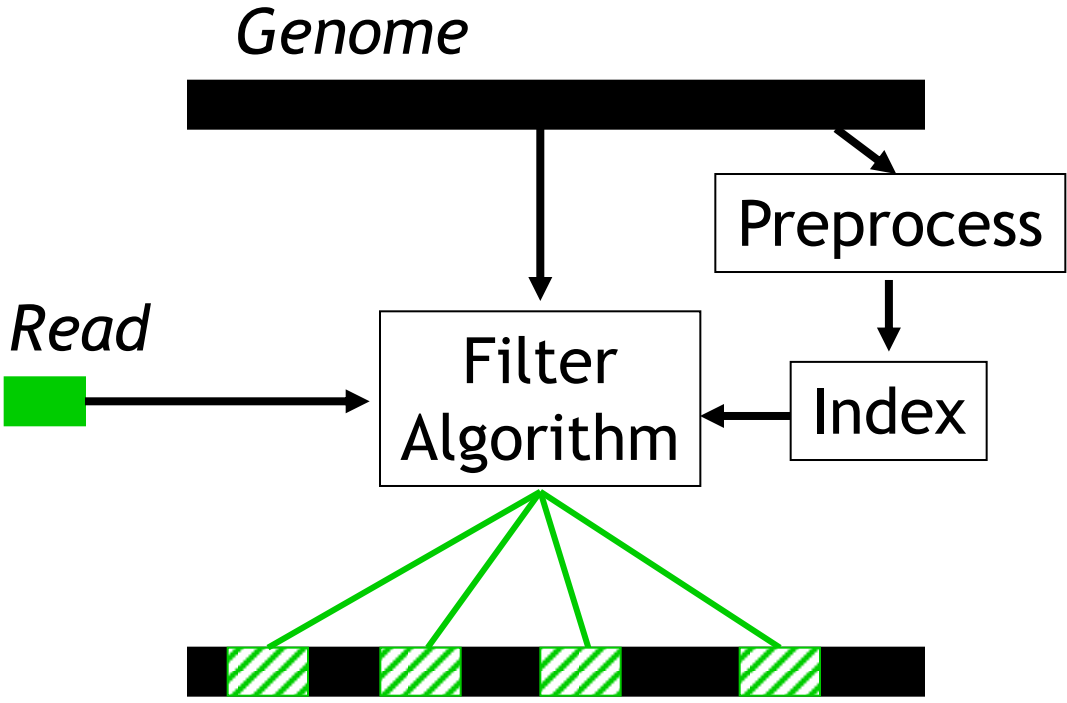


Preprocess

Index



Filtering



Filtration Phase

Potential Matches 

Filtering

Genome



Preprocess

Index

Filter Algorithm

Read



Exact Algorithm



Filtration Phase

Potential Matches 

Verification Phase

True Matches 

False Matches 

Simple k-mer Index, k=3

S = ACGAAAAC TCGATTACTCGACC

	Hitlist		Hitlist		Hitlist
AAA		ACC		CGA	
AAC		ACG		...	
AAG		ACT		GAA	
AAT		AGA		...	
ACA		...		TTT	

- Size of that table: $4^3 = 64$ entries = $|\Sigma|^k$

Simple k-mer Index, k=3

S = **ACG**AAAAC TCGATTACTCGACC

	Hitlist		Hitlist		Hitlist
AAA		ACC		CGA	
AAC		ACG	0	...	
AAG		ACT		GAA	
AAT		AGA		...	
ACA		...		TTT	

- Size of that table: $4^3 = 64$ entries = $|\Sigma|^k$

Simple k-mer Index, k=3

S = A**CGA**AAACTCGATTACTCGACC

	Hitlist		Hitlist		Hitlist
AAA		ACC		CGA	1
AAC		ACG	0	...	
AAG		ACT		GAA	
AAT		AGA		...	
ACA		...		TTT	

- Size of that table: $4^3 = 64$ entries = $|\Sigma|^k$

Simple k-mer Index, k=3

S = AC**GAA**AACTCGATTACTCGACC

	Hitlist		Hitlist		Hitlist
AAA		ACC		CGA	1
AAC		ACG	0	...	
AAG		ACT		GAA	2
AAT		AGA		...	
ACA		...		TTT	

- Size of that table: $4^3 = 64$ entries = $|\Sigma|^k$

Simple k-mer Index, k=3

S = ACGAAAAC TCGATTACTCGACC

	Hitlist		Hitlist		Hitlist
AAA	3,4	ACC	19	CGA	1
AAC	5	ACG	0
AAG	Empty	ACT	6,14	GAA	2
AAT	Empty	AGA
ACA	Empty	TTT	Empty

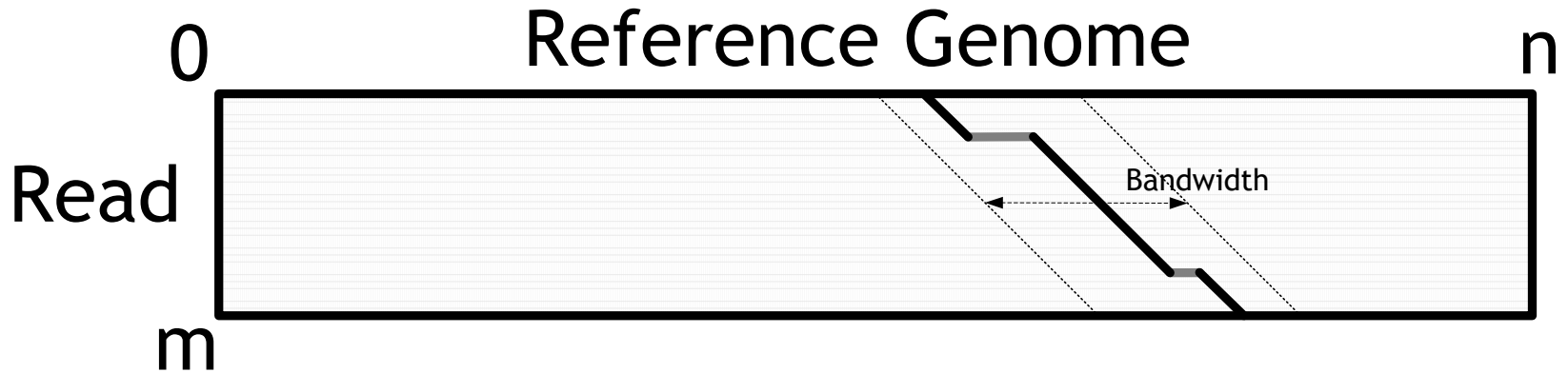
Searching a Read

	Hitlist		Hitlist		Hitlist
AAA	3,4	ACC	19	CGA	1
AAC	5	ACG	0
AAG	Empty	ACT	6,14	GAA	2
AAT	Empty	AGA
ACA	Empty	TTT	Empty

- Read Sequence: **ACTG**
 - Potential match at position 6 and 14

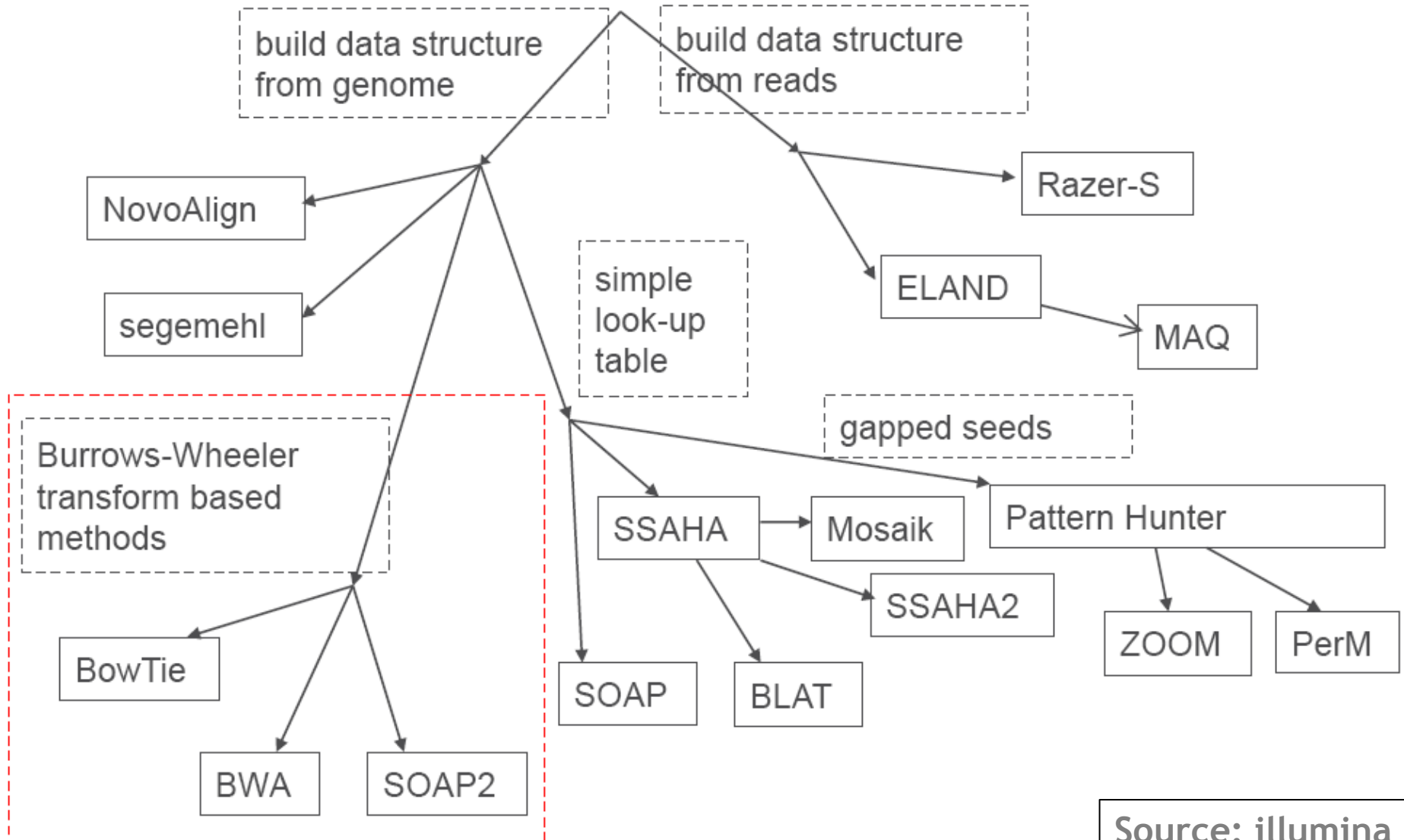
Verification Algorithm

Banded Dynamic Programming

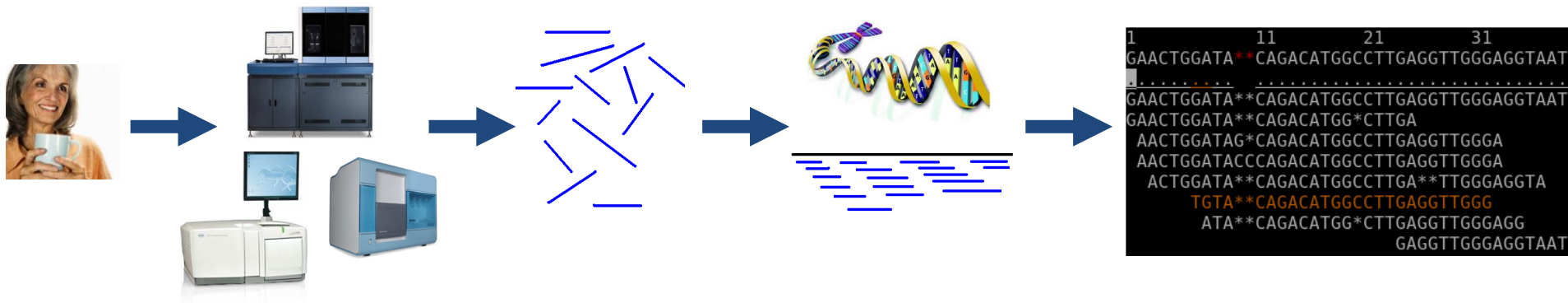




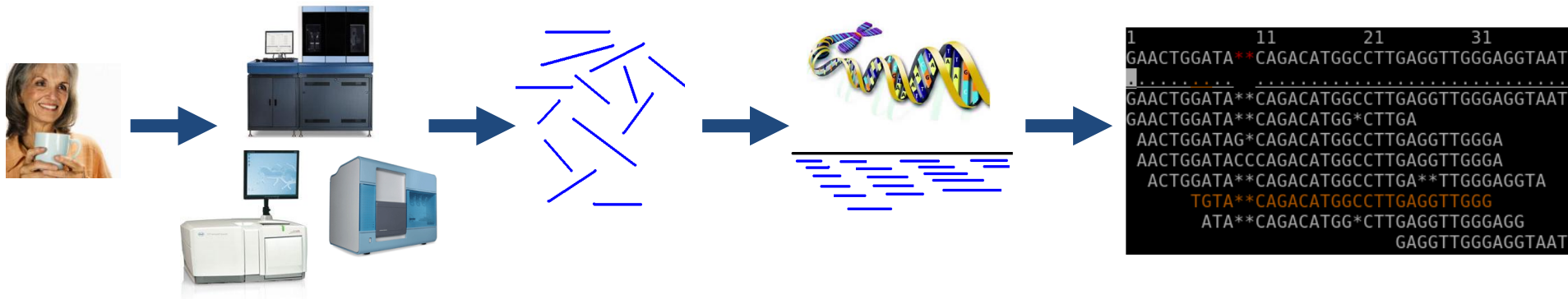
Read Mapping



Genome Capture Analysis

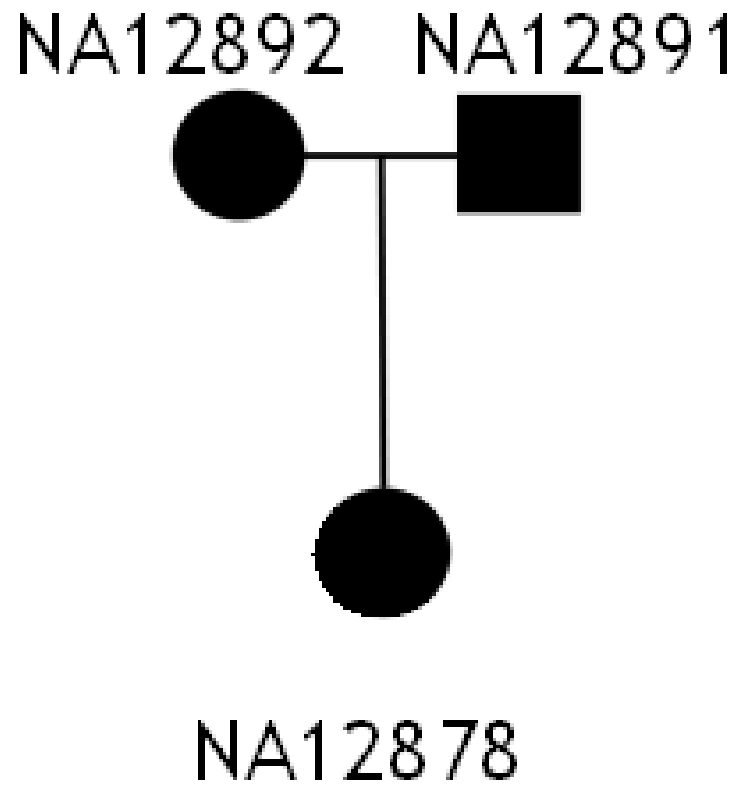


Target Enrichment Analysis

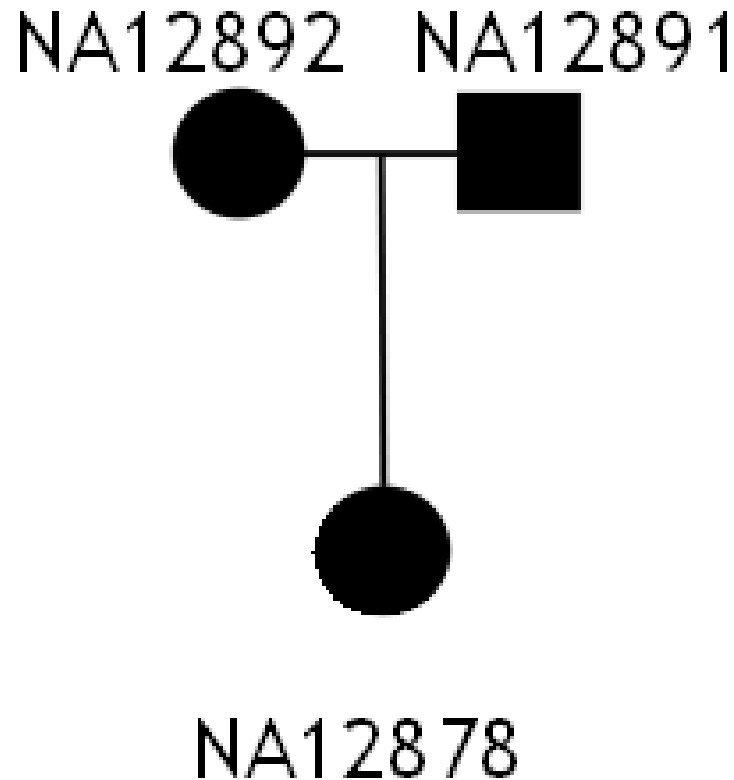


- Quality Control
 - On-target / Off-target Analysis
 - Coverage Analysis
 - GC-Content
- Data Analysis
 - SNP & Short Indel Calling
 - Relating the Variant Calls to Public Databases

HapMap Trio

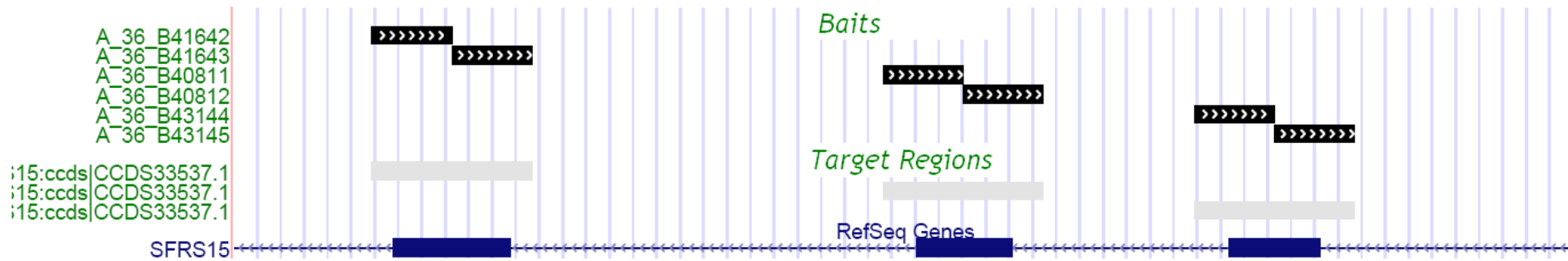


HapMap Trio

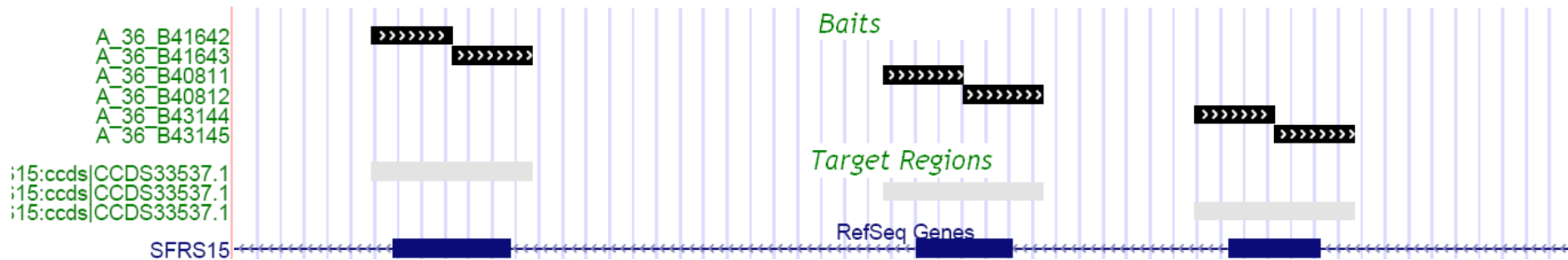


- NA12878 and NA12891 were sequenced

Individual Baits vs. Target Regions

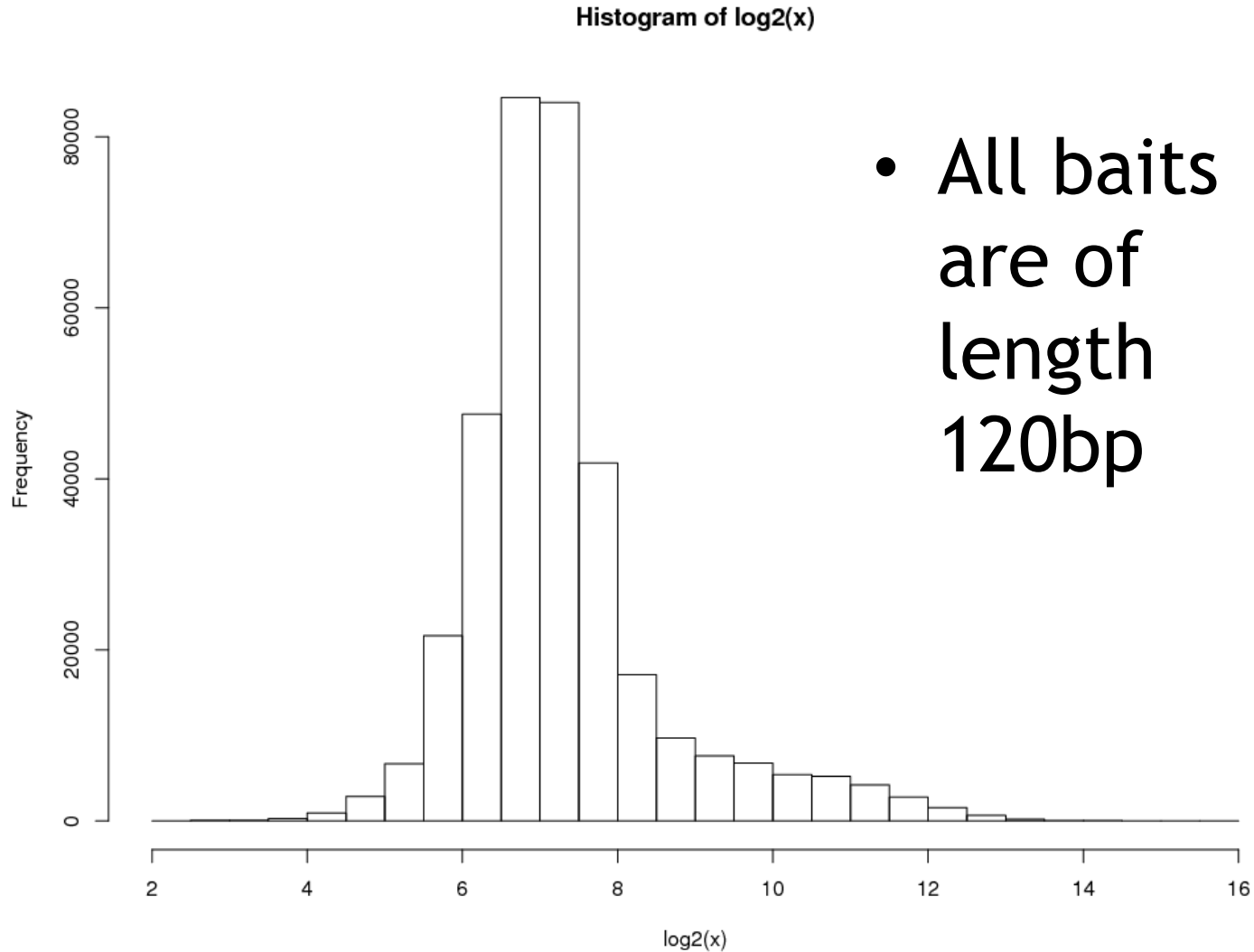


Individual Baits vs. Target Regions



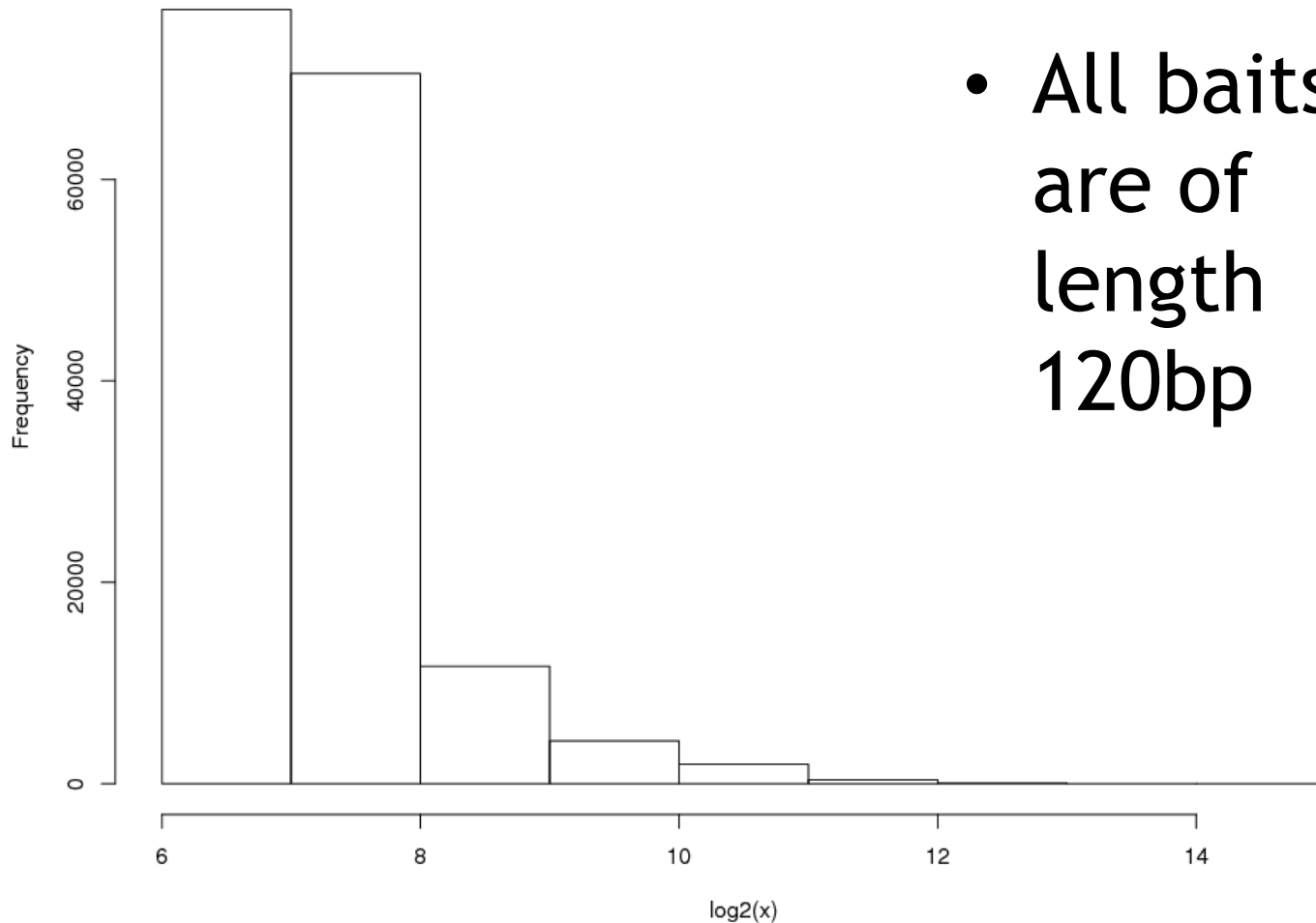
- All bait and target region coordinates are hg18
- Total length of target regions: 37806033 (\approx 38MB)
- Total length of bait sequences: 38235516 (\approx 38MB)
- Approximately 1% of the human genome

Exon Length Distribution



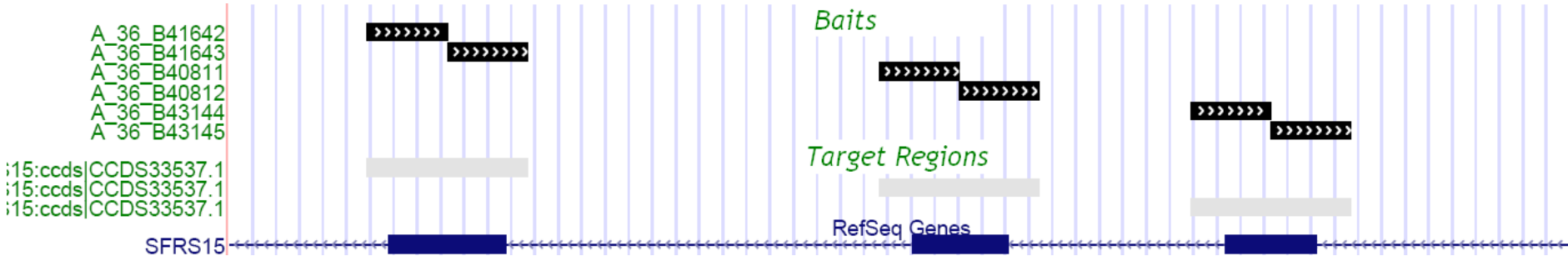
Target Region Length Distribution

Histogram of $\log_2(x)$

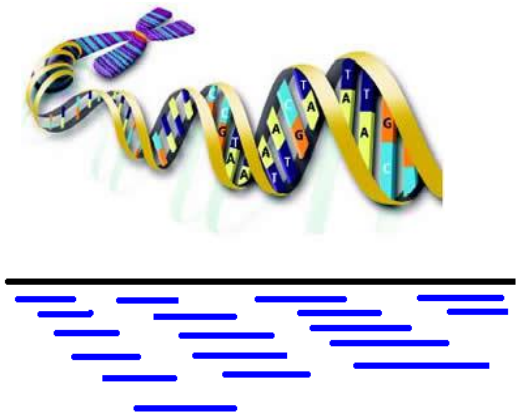


- All baits are of length 120bp

On-target / Off-target Analysis



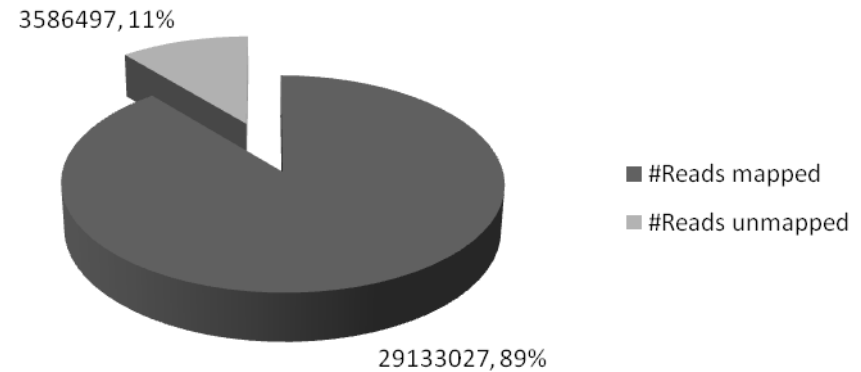
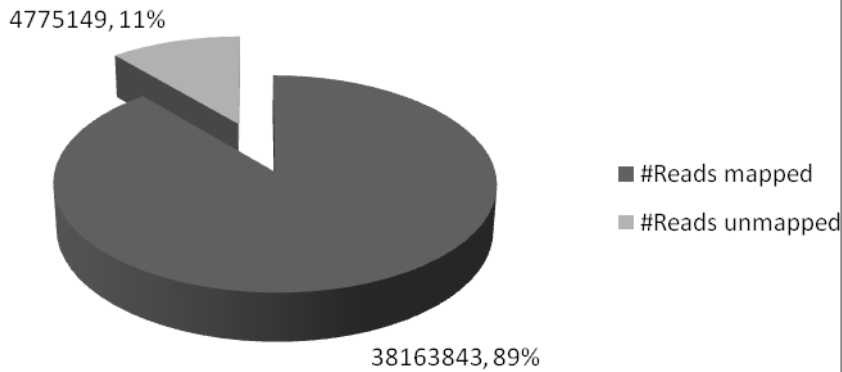
- How many reads are on-target?



Mapped Reads

NA12878

NA12891

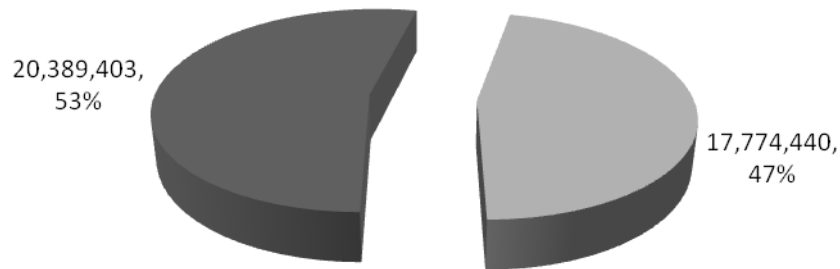


- Mapped SOLiD reads
 - NA12878: 133,915,955 mapped reads
 - NA12891: 108,092,260 mapped reads

On-target / Off-target Analysis

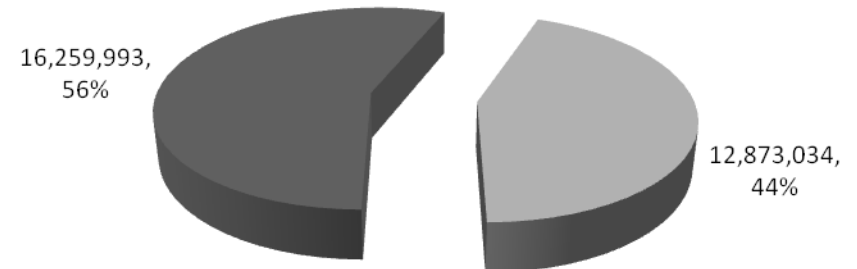
NA12878 - Illumina

■ #Reads in targeted regions ■ #Reads in off-targeted regions



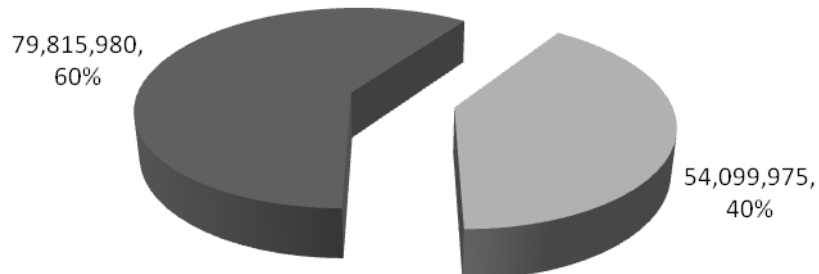
NA12891 - Illumina

■ #Reads in targeted regions ■ #Reads in off-targeted regions



NA12878 - SOLiD

■ #Reads in targeted regions ■ #Reads in off-targeted regions



NA12891 - SOLiD

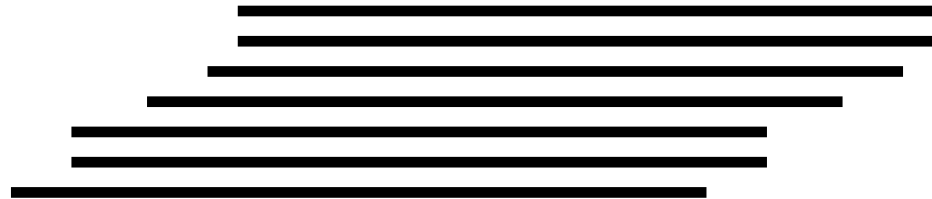
■ #Reads in targeted regions ■ #Reads in off-targeted regions



Duplicates



?

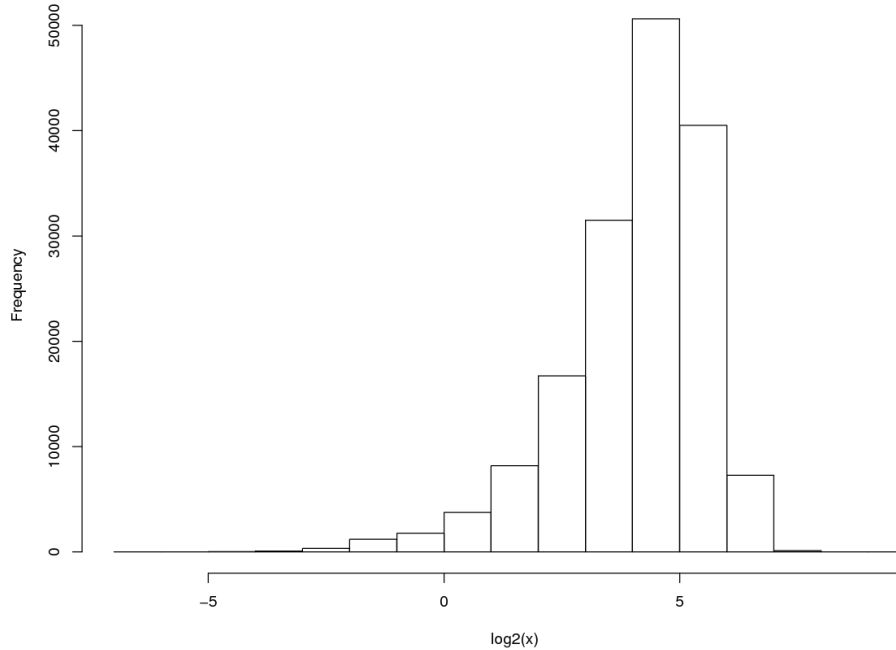


Duplicates

- The Illumina data is paired-end data
 - Redundancy easy to calculate
 - NA12878: 4%
 - NA12891: 5%
 - All redundant Illumina read pairs have been removed
 - For the single-end SOLiD data we did NOT filter the alignments

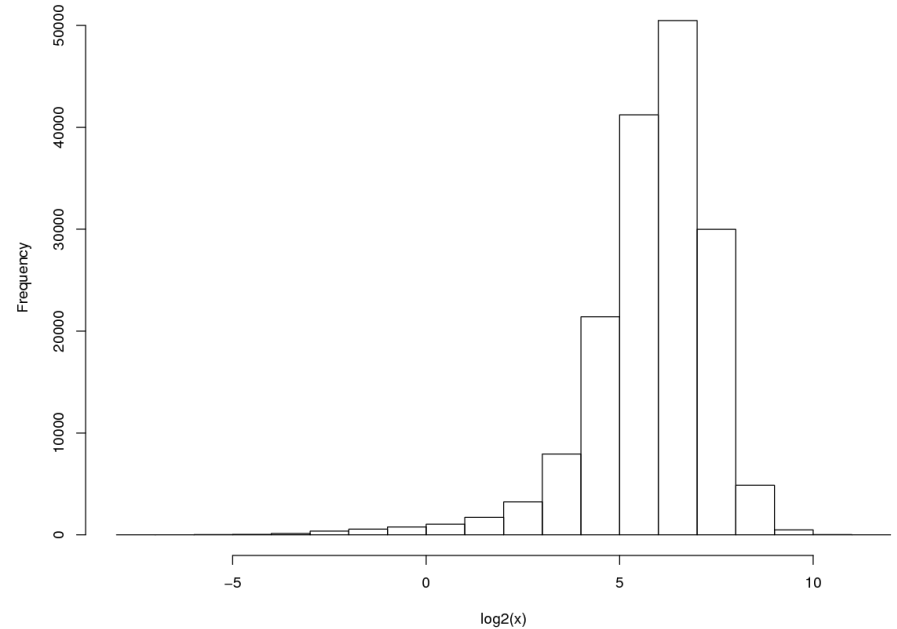
Avg. Coverage for each Target

Histogram of Avg. Coverage – NA12878



Illumina

Histogram of Avg. Coverage – NA12878

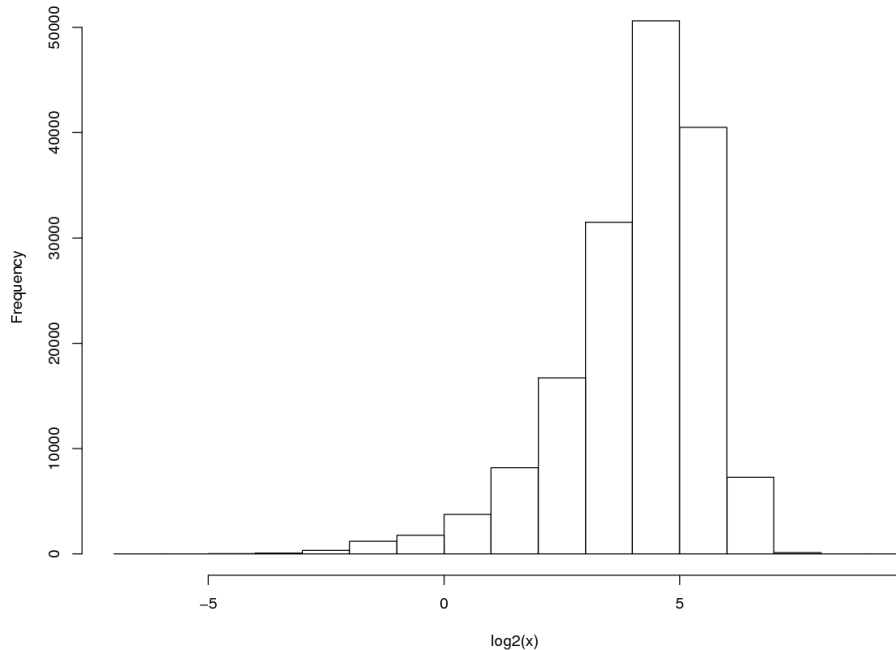


SOLiD

- SOLiD distribution is shifted to the right due to higher sequencing coverage

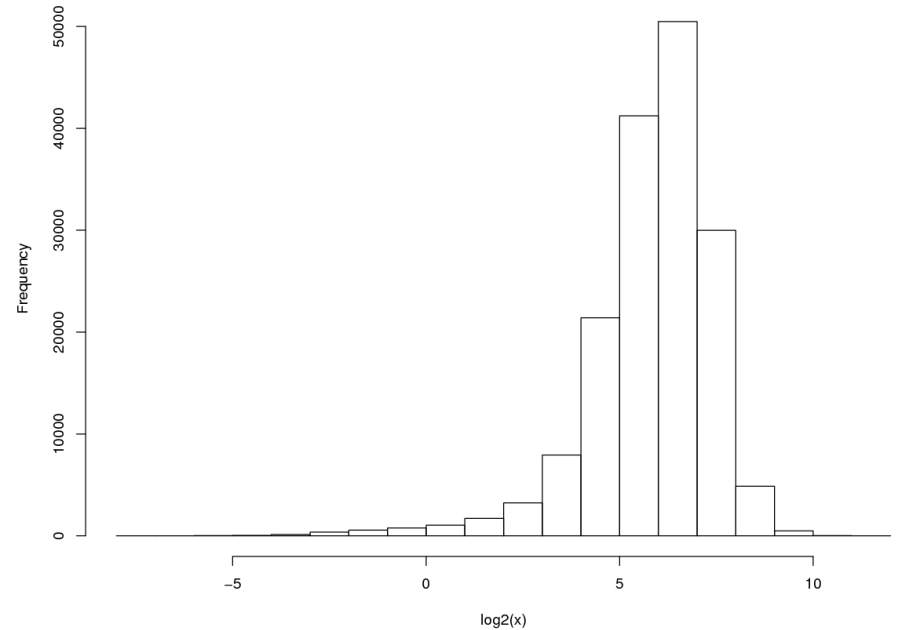
Avg. Coverage for each Target

Histogram of Avg. Coverage – NA12878



- 3490 Targets without any mapped base

Histogram of Avg. Coverage – NA12878

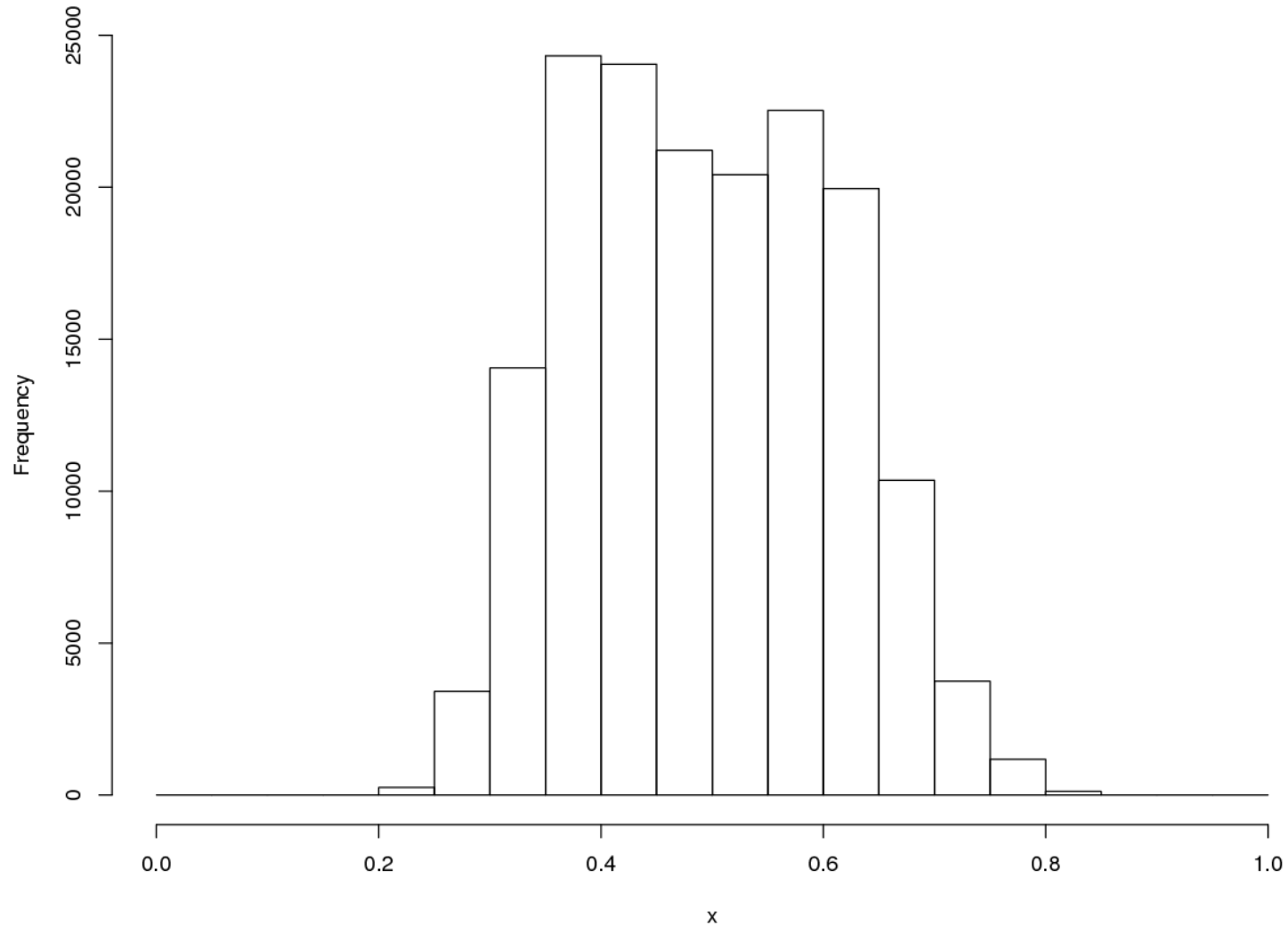


- 1280 Targets without any mapped base

← Overlap: 825 →

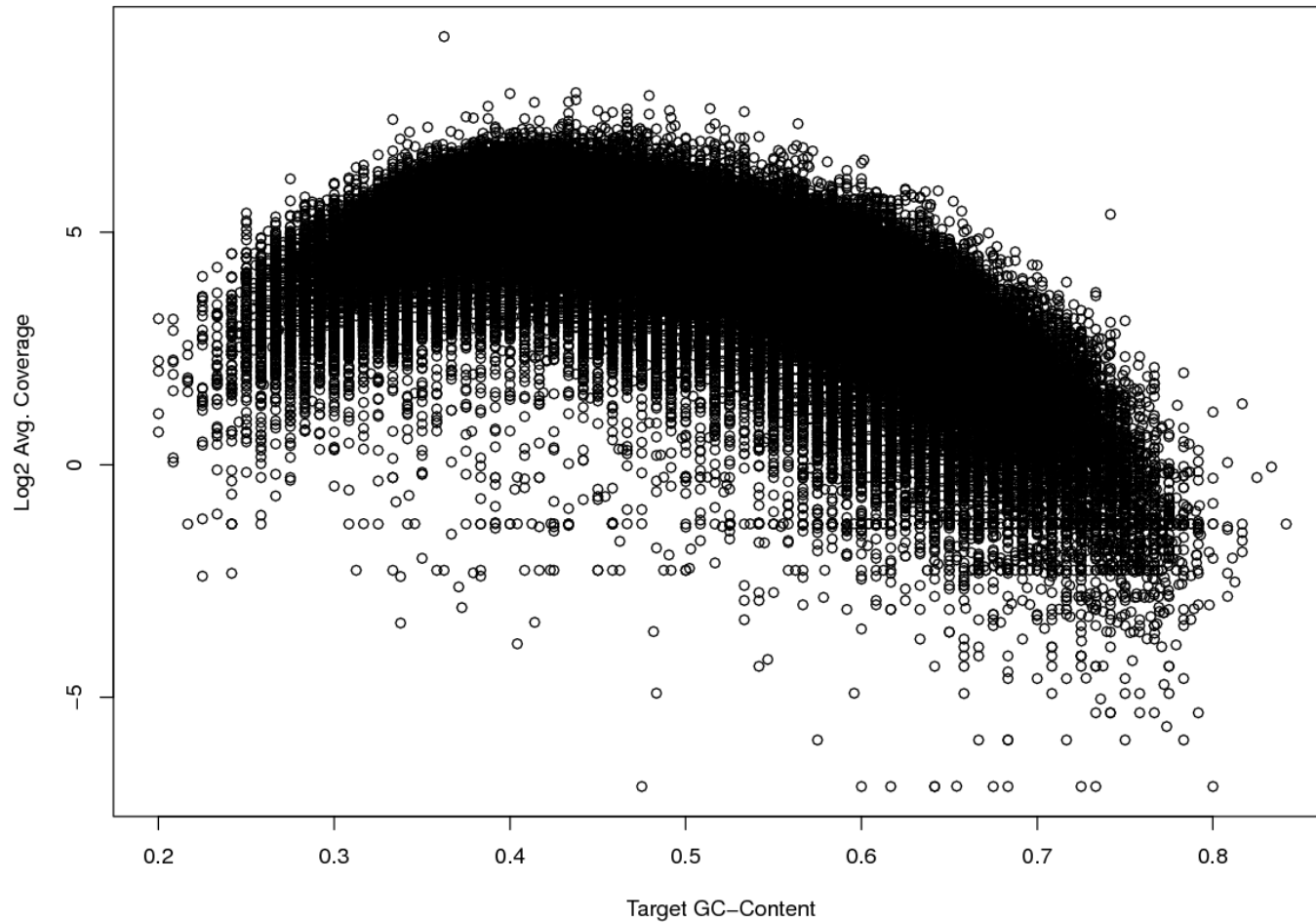
GC-Content Distribution

Target GC-Content Distribution



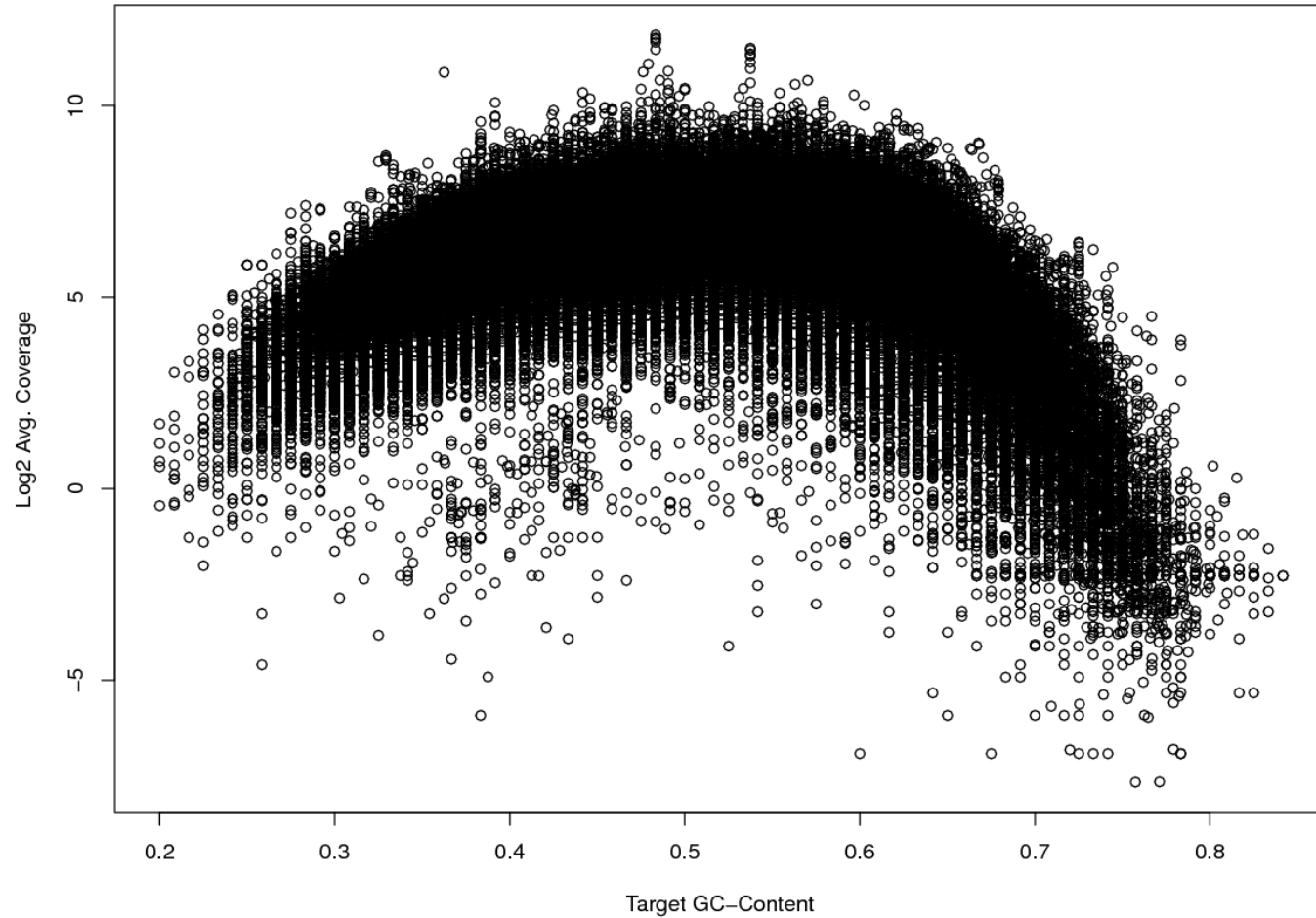
GC-Content Distribution

NA12878 – Illumina

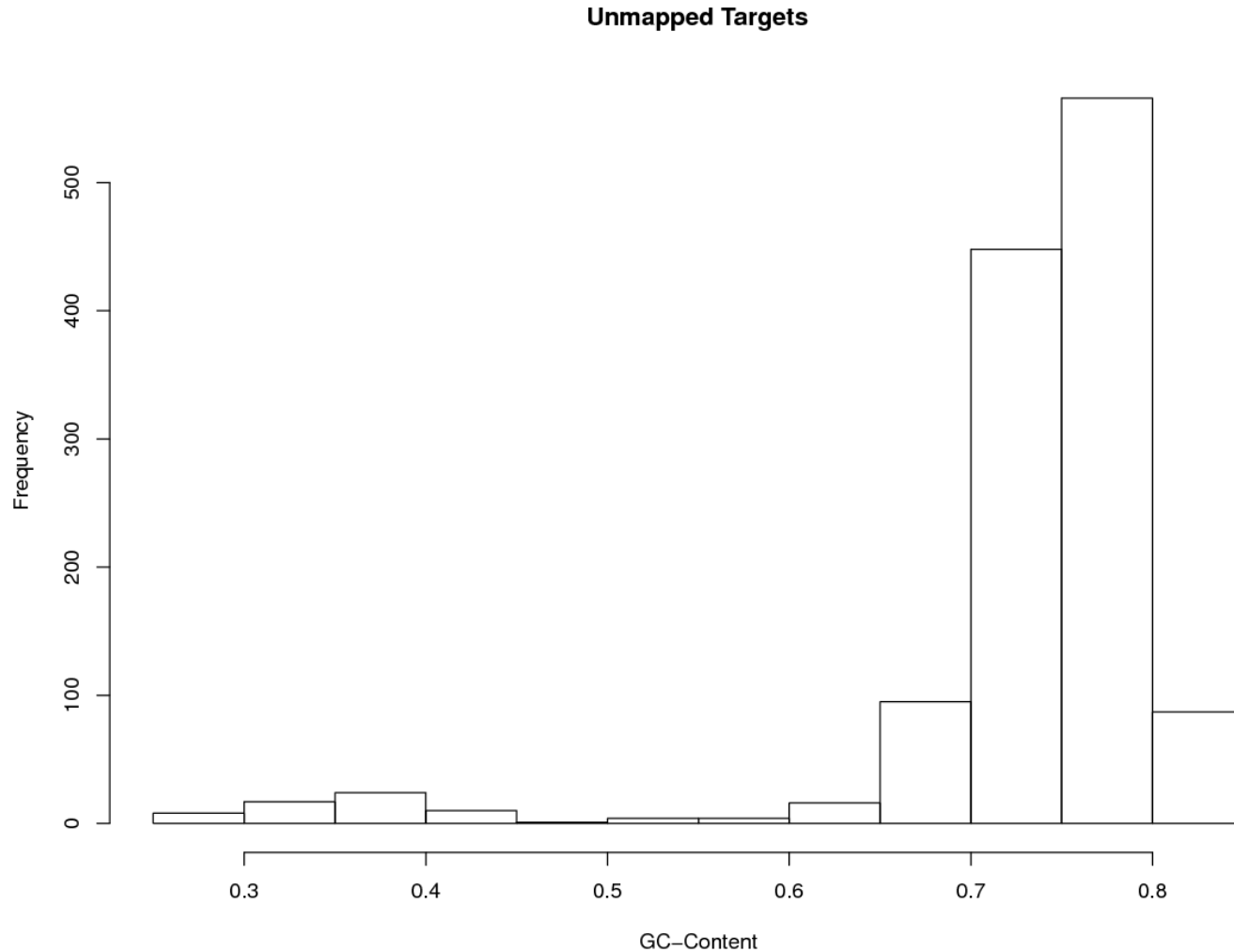


GC-Content Distribution

NA12878 – SOLiD

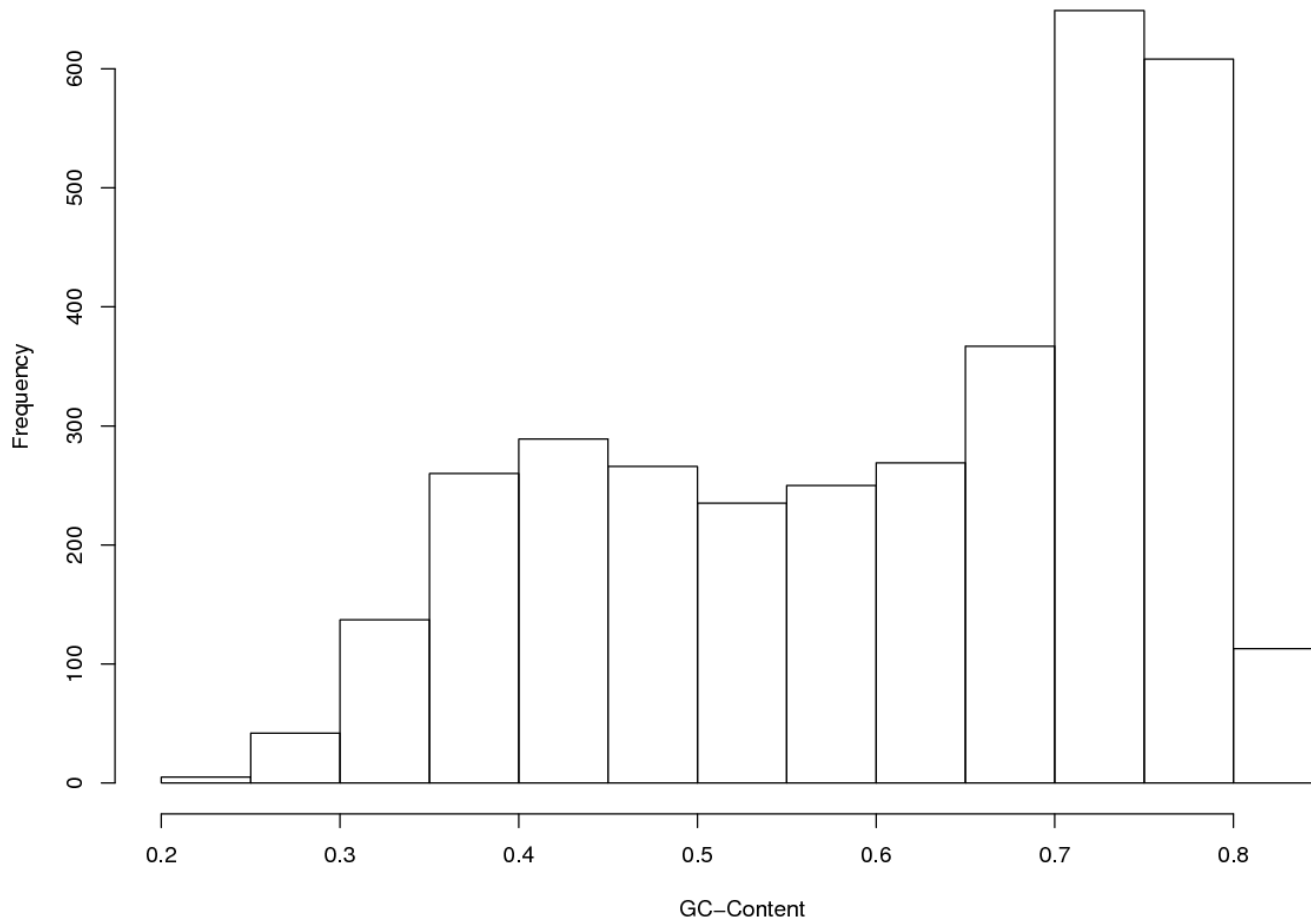


Histogram of GC-Content of Unmapped Targets - SOLiD

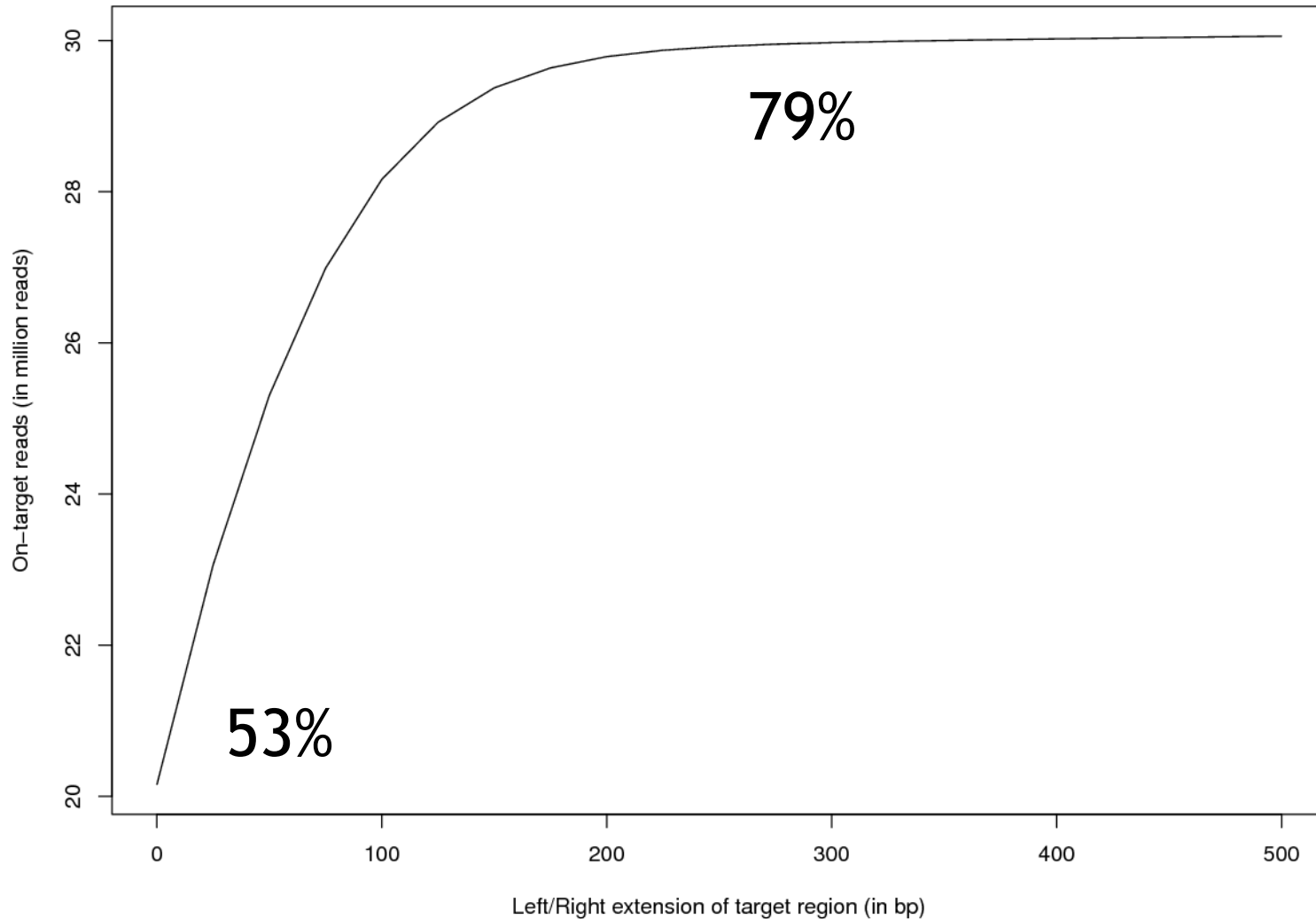


Histogram of GC-Content of Unmapped Targets - Illumina

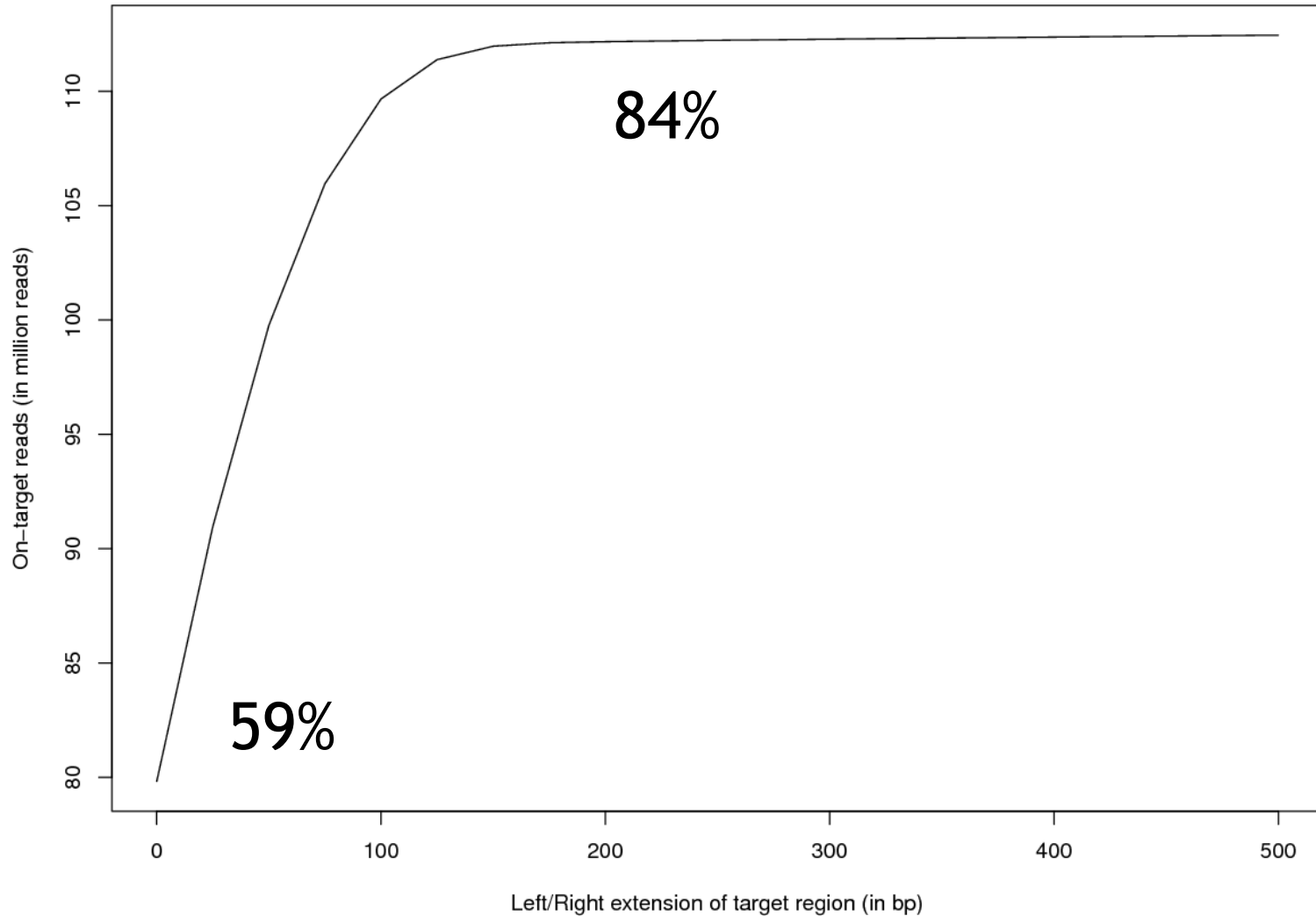
Unmapped Targets



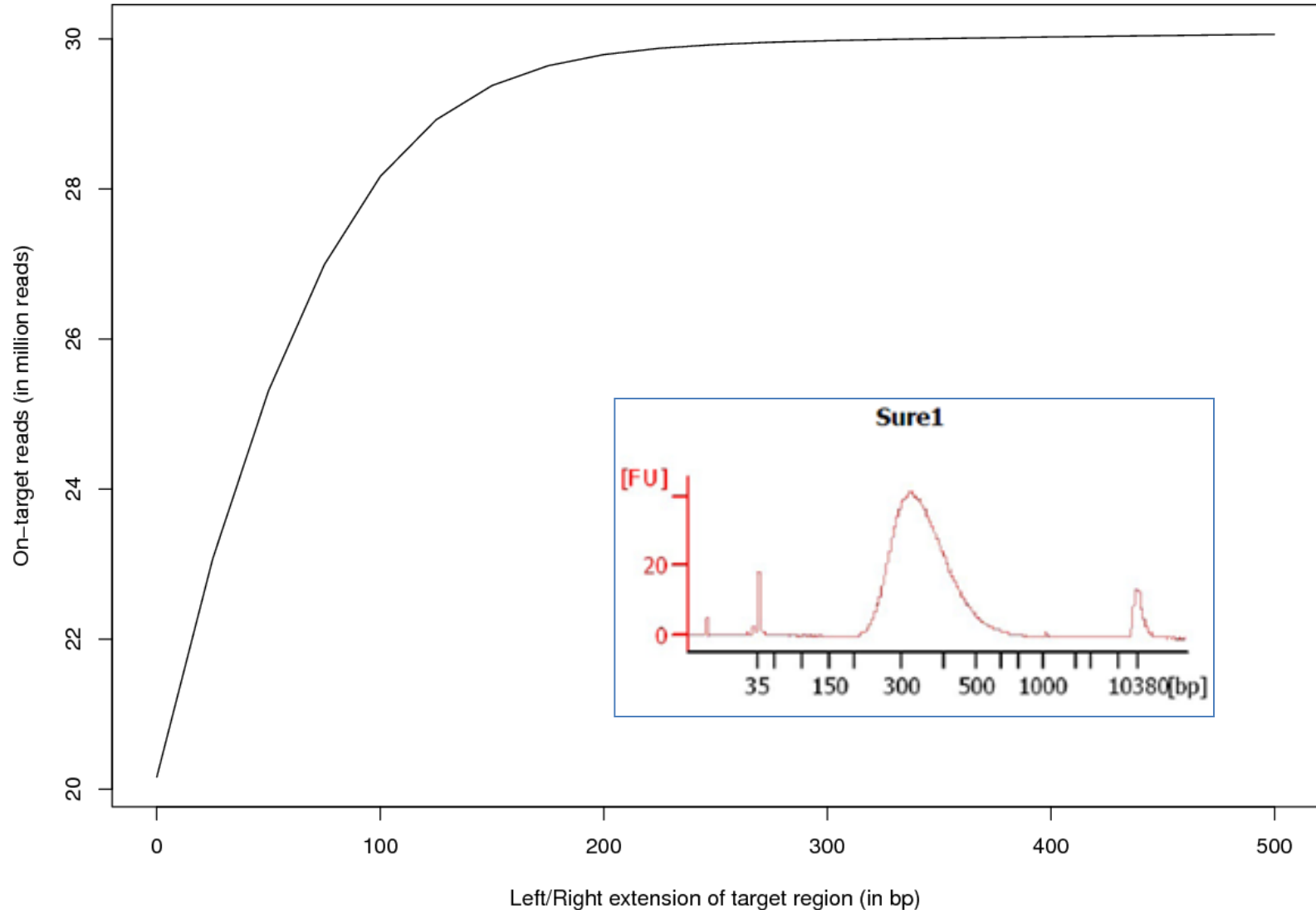
Where did the off-target reads end-up?



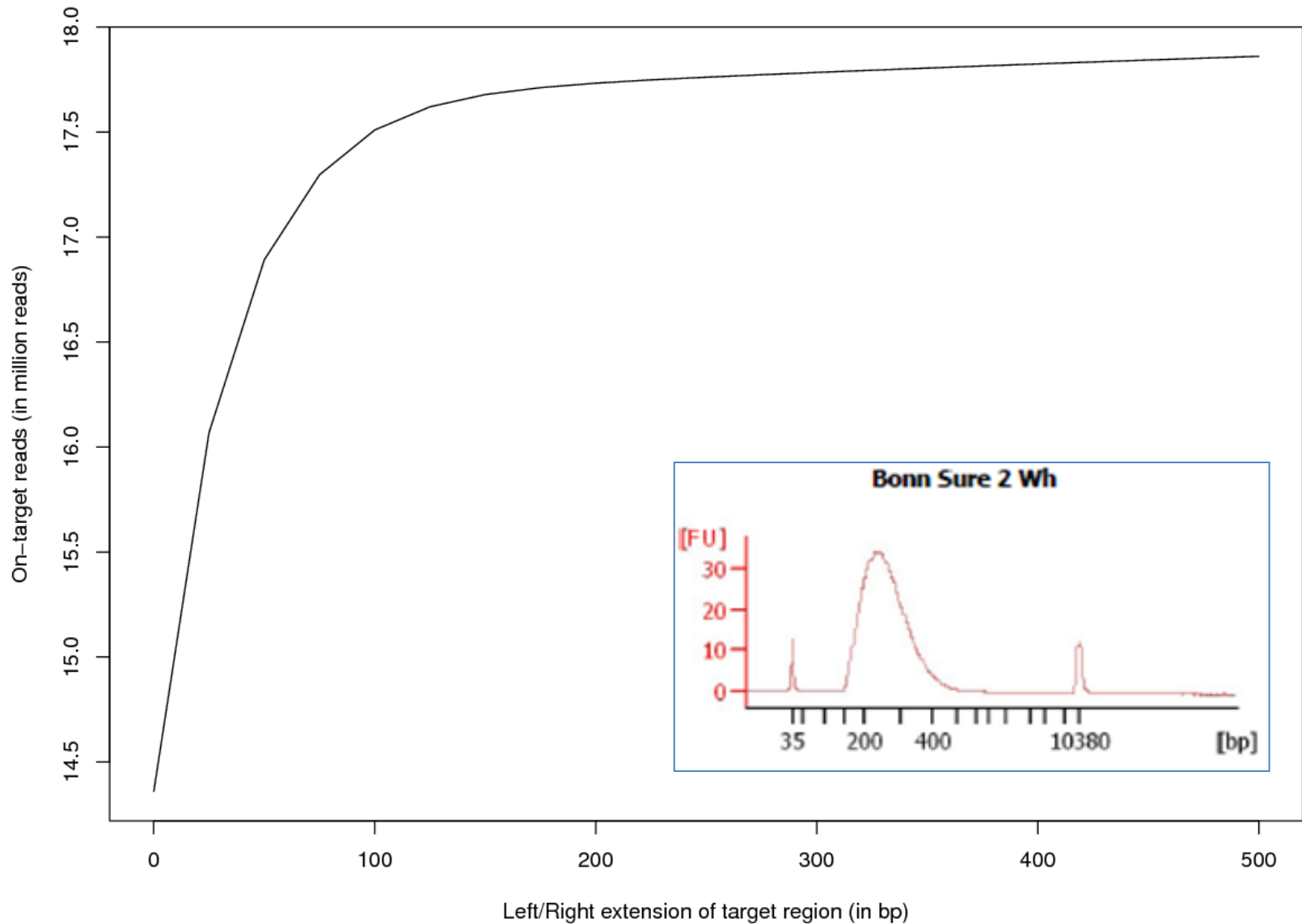
Where did the off-target reads end-up?



Insert Size is very important!



Different Sample: Smaller insert size

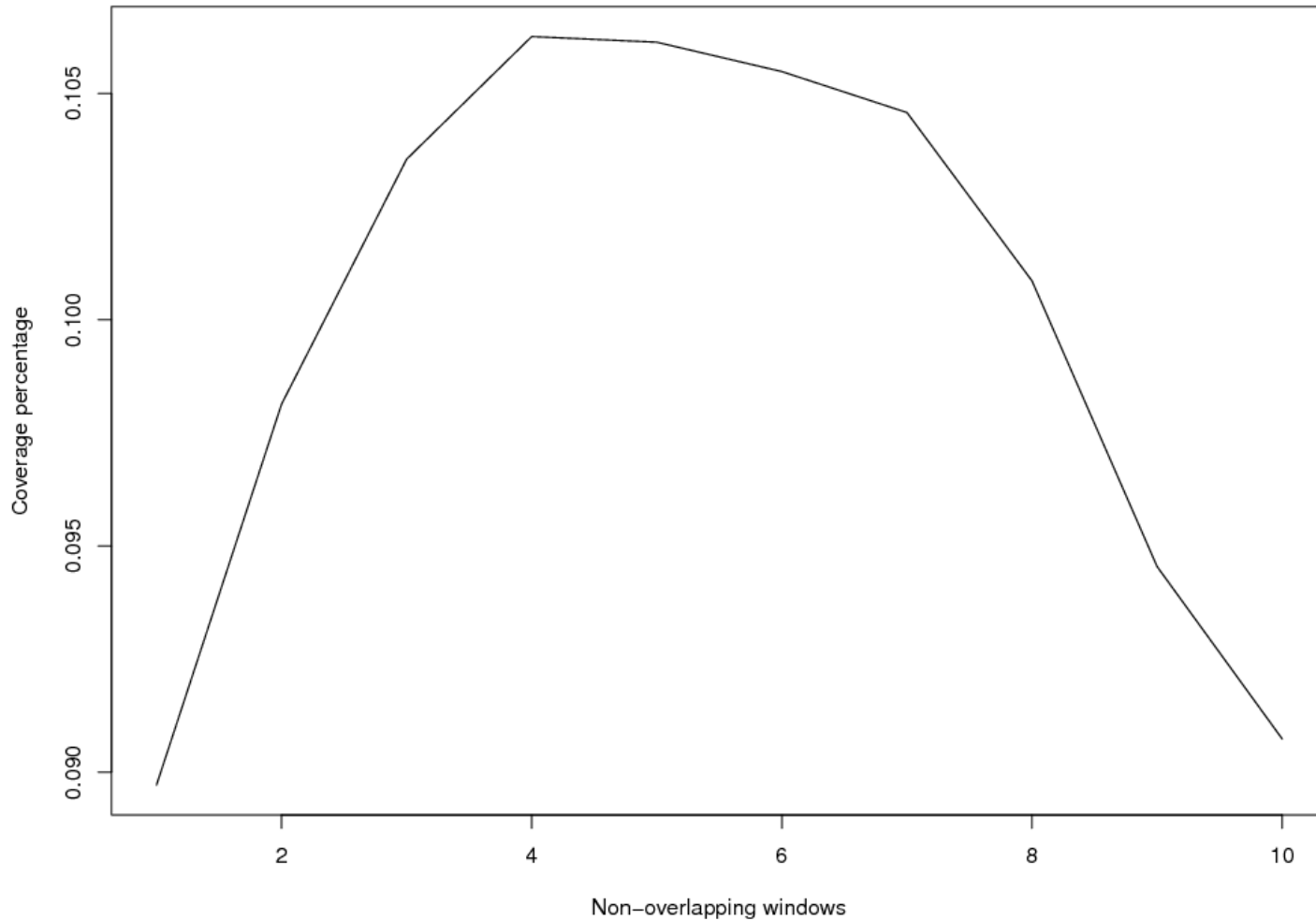


Uniform Coverage across Targets?

- Subdivide each target into 10 non-overlapping windows
- Calculate coverage for each window
- Get the fractional coverage for each window compared to the total coverage of the target
- Add up all fractions for the same window across all targets

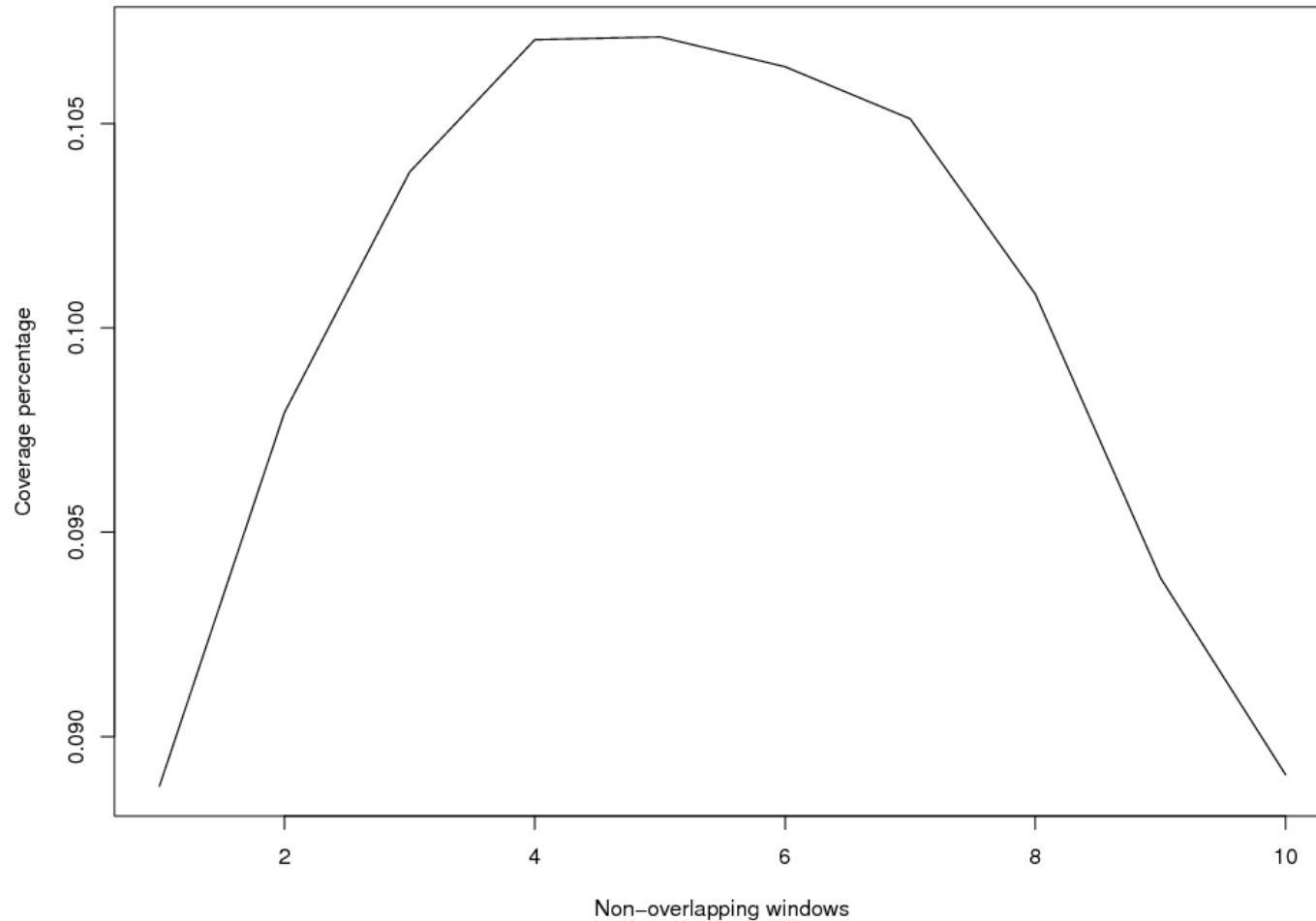
Uniform Coverage across Targets?

NA12878 - Illumina

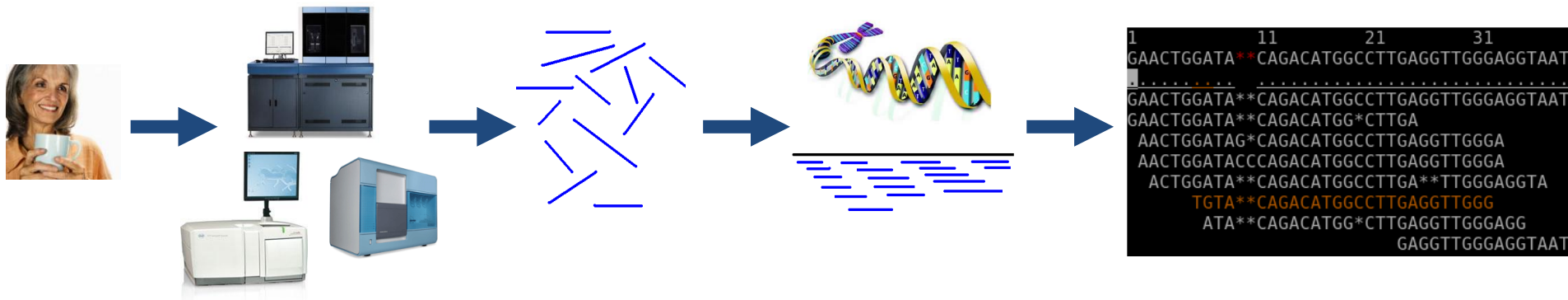


Uniform Coverage across Targets?

NA12891 - Illumina

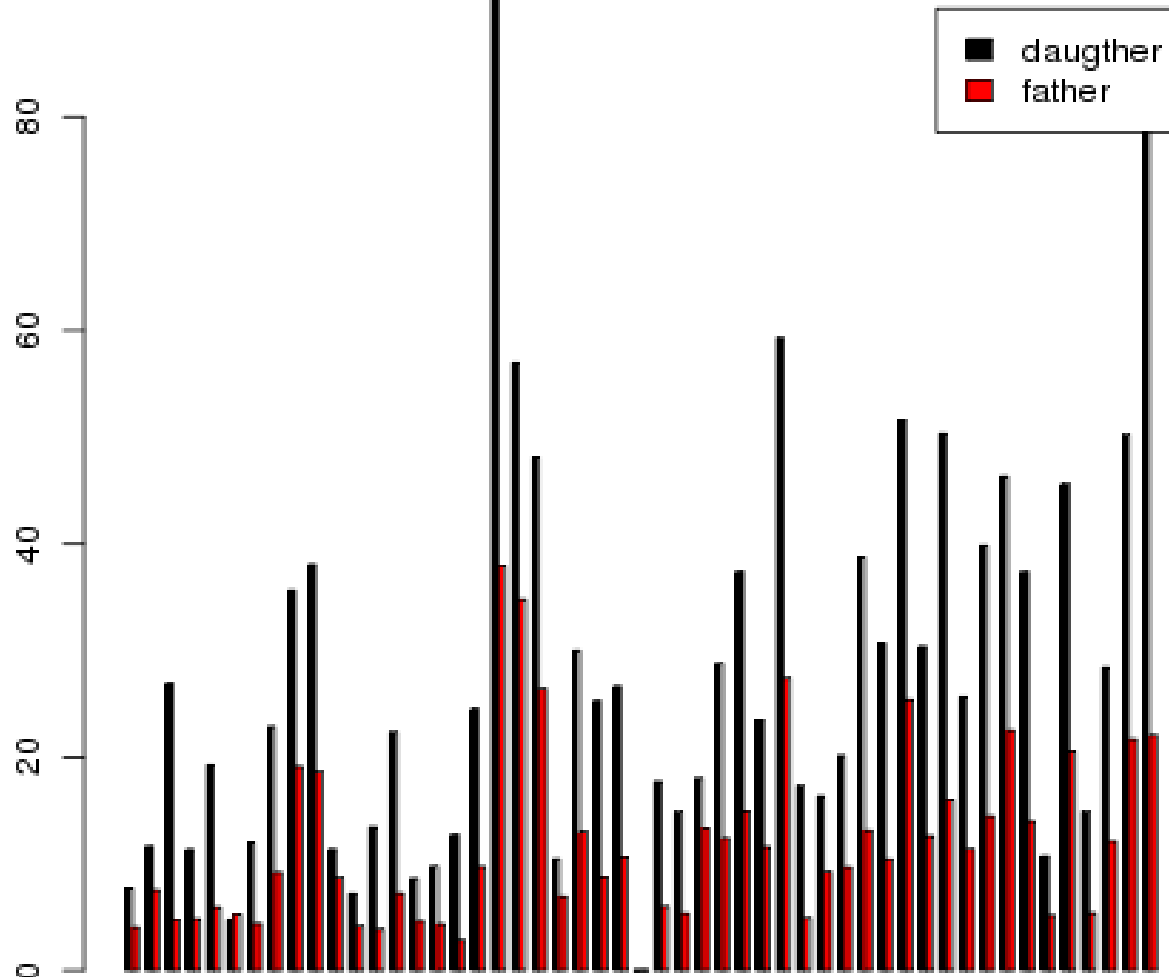


Genome Capture Analysis



- Downstream Analysis
 - chrX and chrY
 - SNP & Short Indel Calling
 - Relating the Variant Calls to Public Databases

50 random targets on chrX



Fraction of Reads on chrX

- Illumina data
 - NA12878: 0.02475
 - NA12891: 0.01329
- SOLiD data
 - NA12878: 0.02995
 - NA12891: 0.01162

Read Counts on chrY

- Illumina data
 - NA12878: 85 reads
 - NA12891: 14953 reads
- SOLiD data
 - NA12878: 17890 reads
 - NA12891: 125638 reads

SNP Calling



dbSNP

1000 Genomes
A Deep Catalog of Human Genetic Variation

Individual
Genotype
Information

SNP Calls

```
1      11      21      31
GAACTGATA**CAGACATGGCCTTGAGGTTGGGAGGTAAT
.....
GAACTGATA**CAGACATGGCCTTGAGGTTGGGAGGTAAT
GAACTGATA**CAGACATGG*CTTGA
AACTGGATAG*CAGACATGGCCTTGAGGTTGGGA
AACTGGATACCCAGACATGGCCTTGAGGTTGGGA
ACTGGATA**CAGACATGGCCTTGA**TTGGGAGGTA
TGTA**CAGACATGGCCTTGAGGTTGGG
ATA**CAGACATGG*CTTGAGGTTGGGAGG
GAGGTTGGGAGGTAAT
```

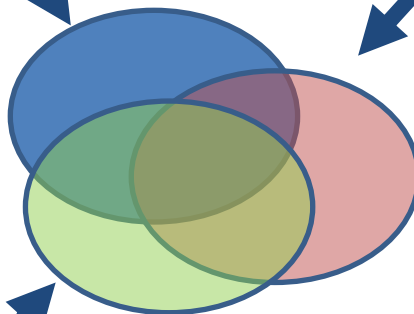
Short Indel Calling



dbSNP

1000 Genomes
A Deep Catalog of Human Genetic Variation

Individual
Genotype
Information

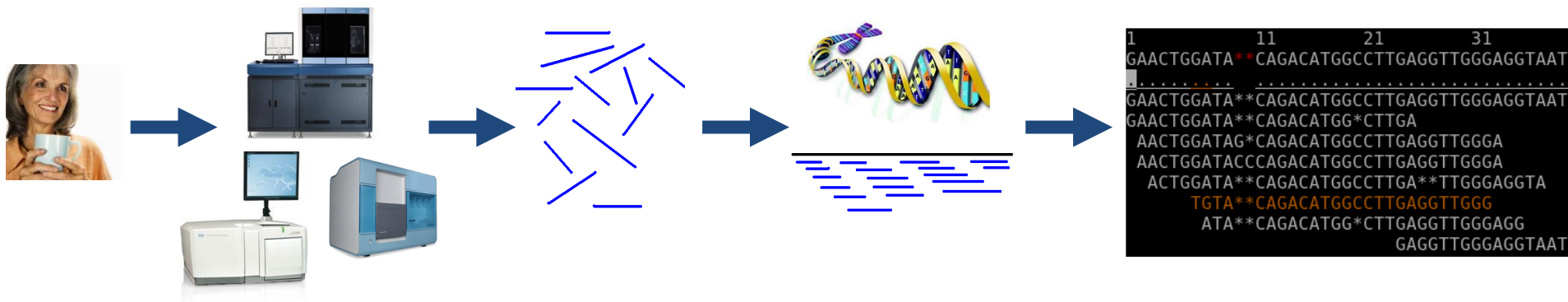


Indel Calls

```
1      11      21      31
GAACTGGATA**CAGACATGGCCTTGAGGTTGGGAGGTAAT
.....
GAACTGGATA**CAGACATGGCCTTGAGGTTGGGAGGTAAT
GAACTGGATA**CAGACATGG*CTTGA
AACTGGATAG*CAGACATGGCCTTGAGGTTGGGA
AACTGGATACCCAGACATGGCCTTGAGGTTGGGA
ACTGGATA**CAGACATGGCCTTGA**TTGGGAGGTA
TGTA**CAGACATGGCCTTGAGGTTGGG
ATA**CAGACATGG*CTTGAGGTTGGGAGG
GAGGTTGGGAGGTAAT
```

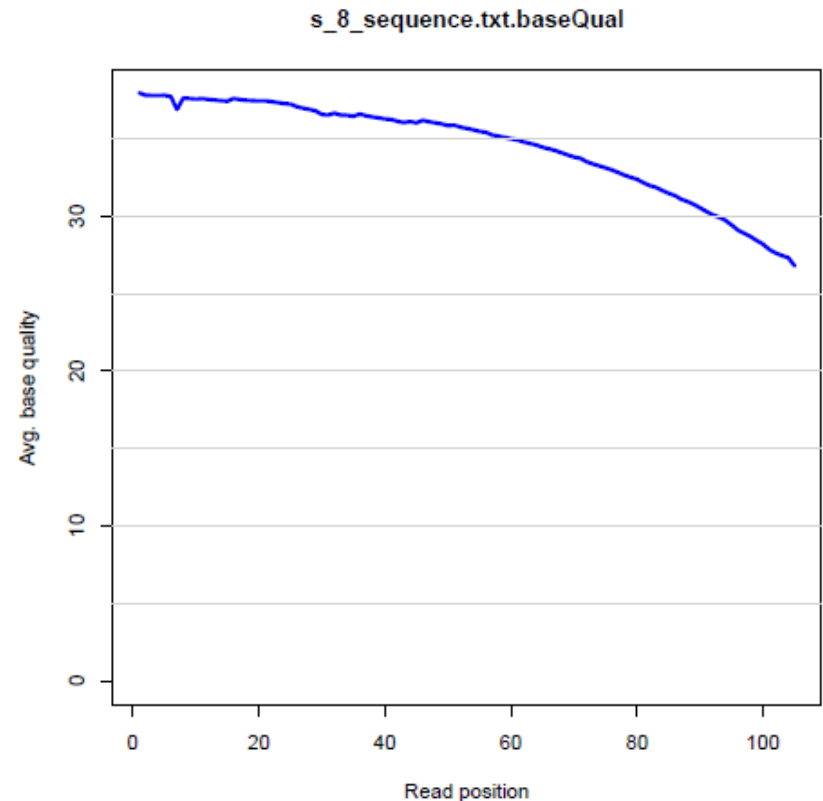
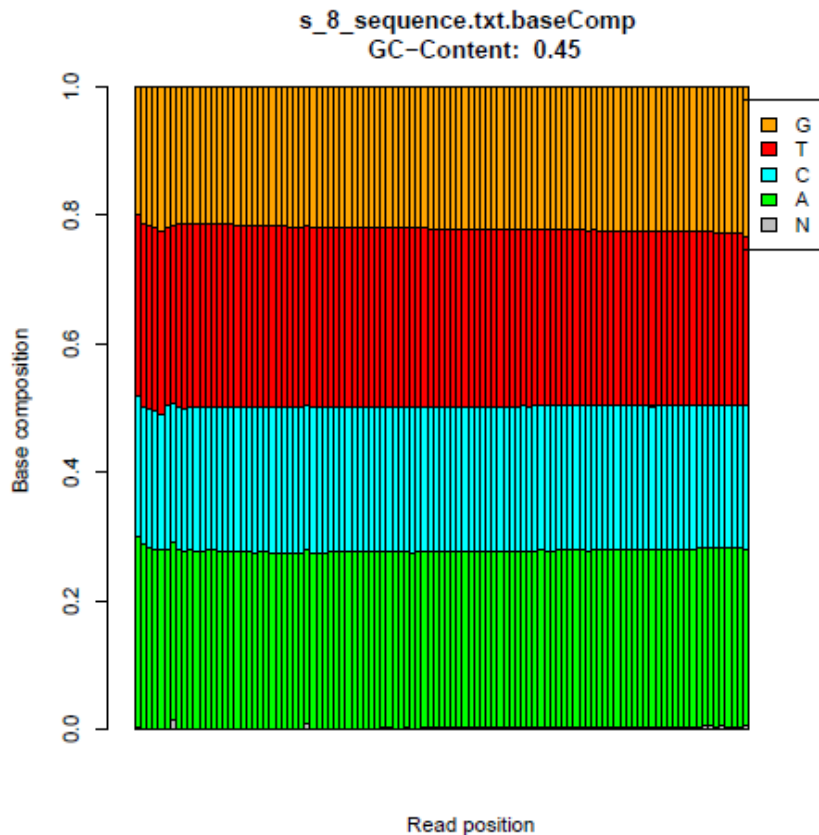
Target Enrichment

Recent Results, May 2011



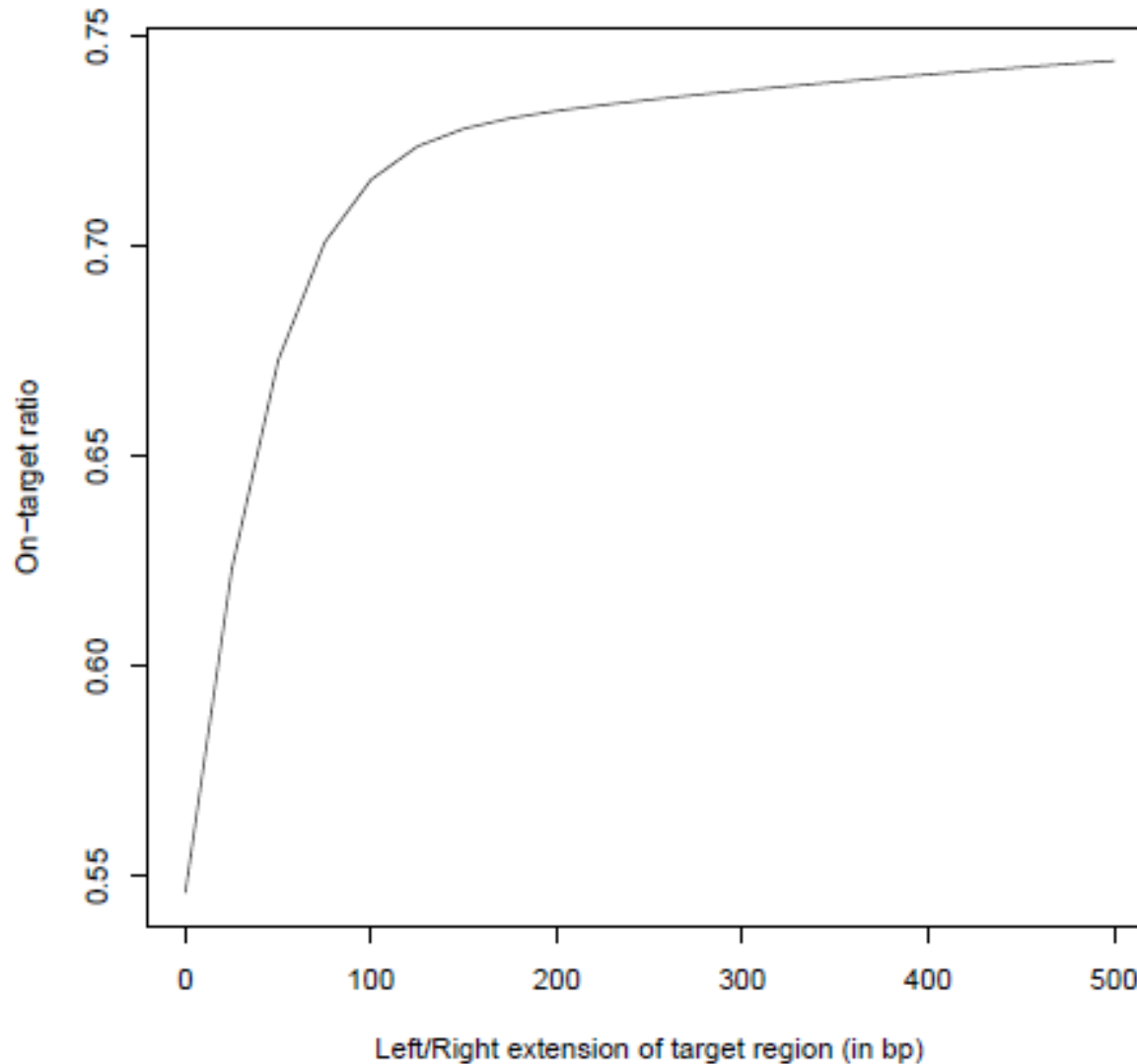
GA Lane, 50MB Kit

- 37 million reads, 32 million mapped (86%)
- 105bp reads



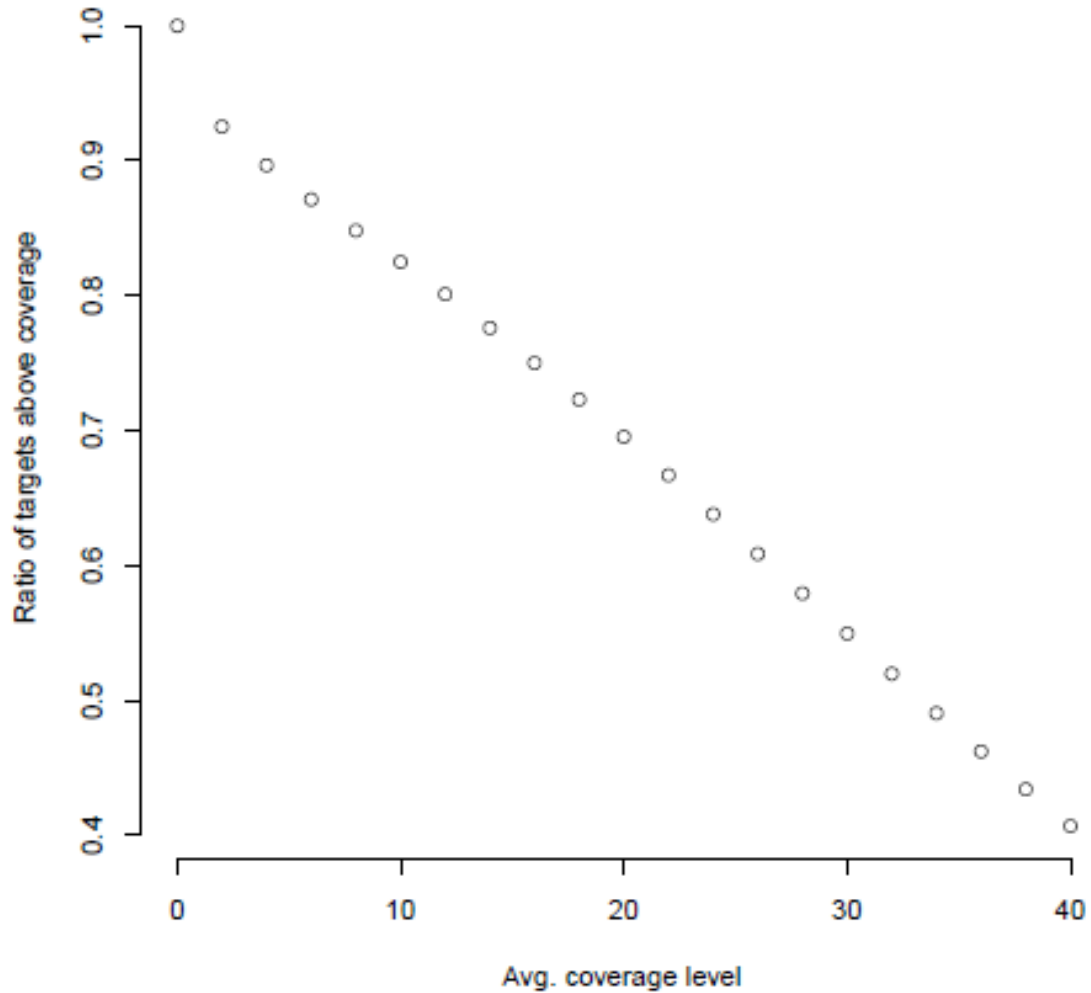
GA Lane, 50MB Kit

s_8_sequence_SureSelect.bed



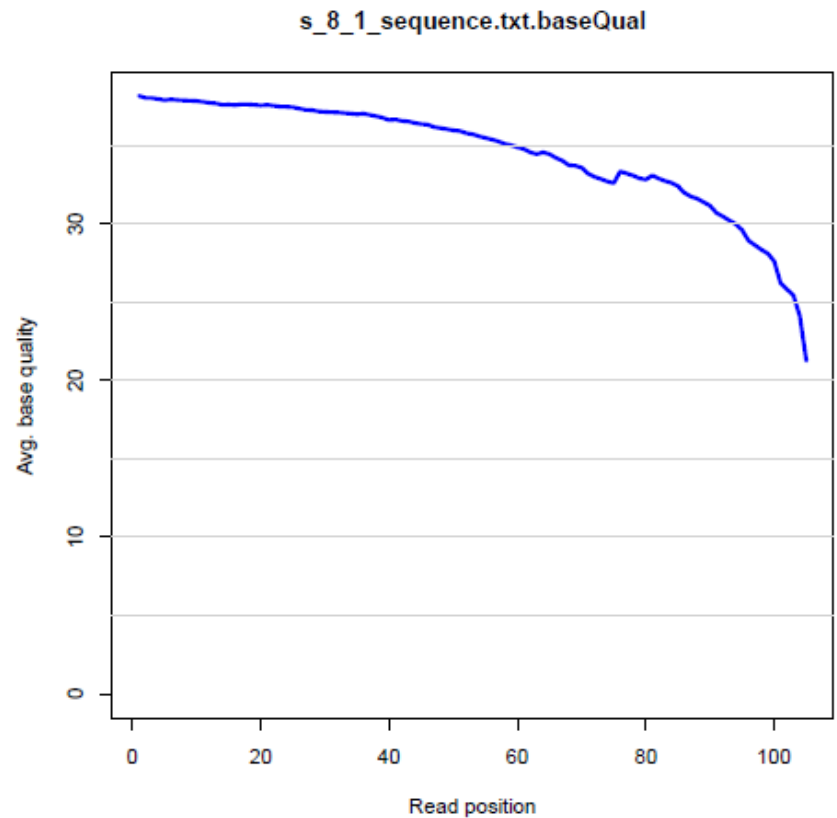
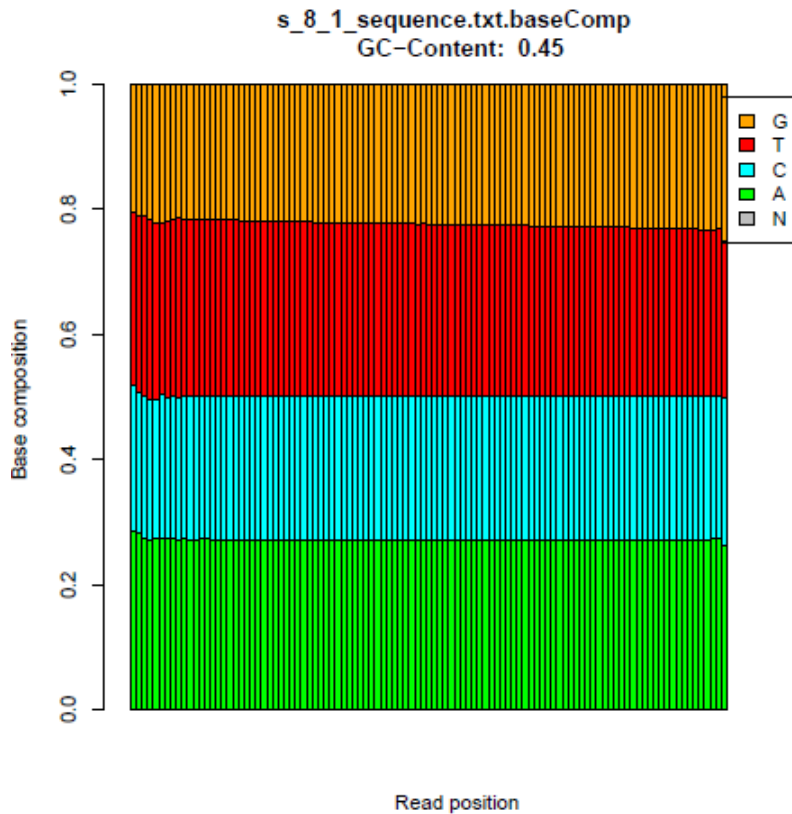
GA Lane, 50MB Kit

s_8_sequence_SureSelect.bed



HiSeq Lane, 50MB Kit

- 185 million reads, 160 million mapped (86%)
- 105bp paired-end reads

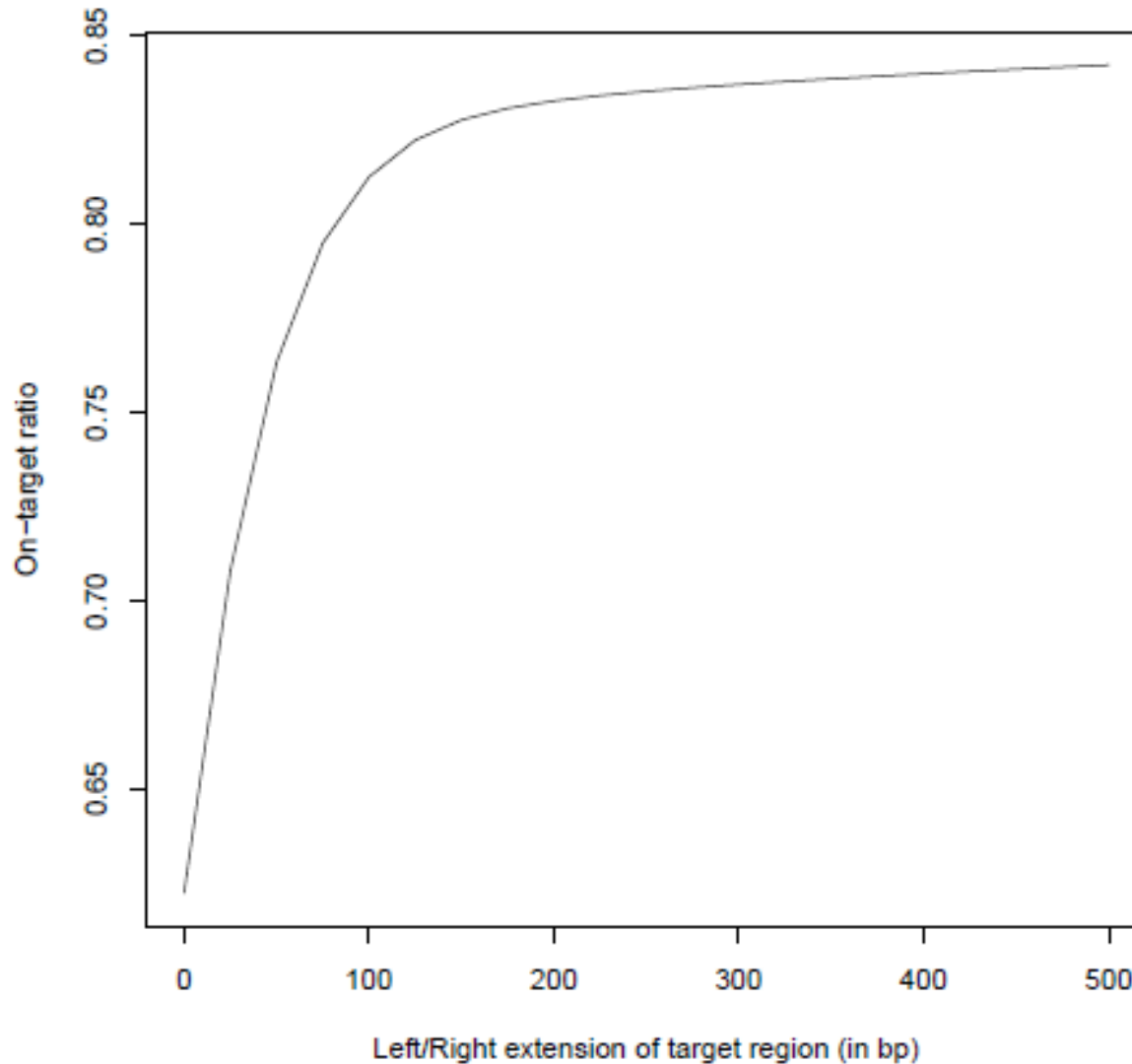


Redundancy

- 185 million reads, 160 million mapped (86%)
- 105bp paired-end reads
- Reads mapped to the same chr: 157 million
- Non-redundant reads: 130 million (82%)

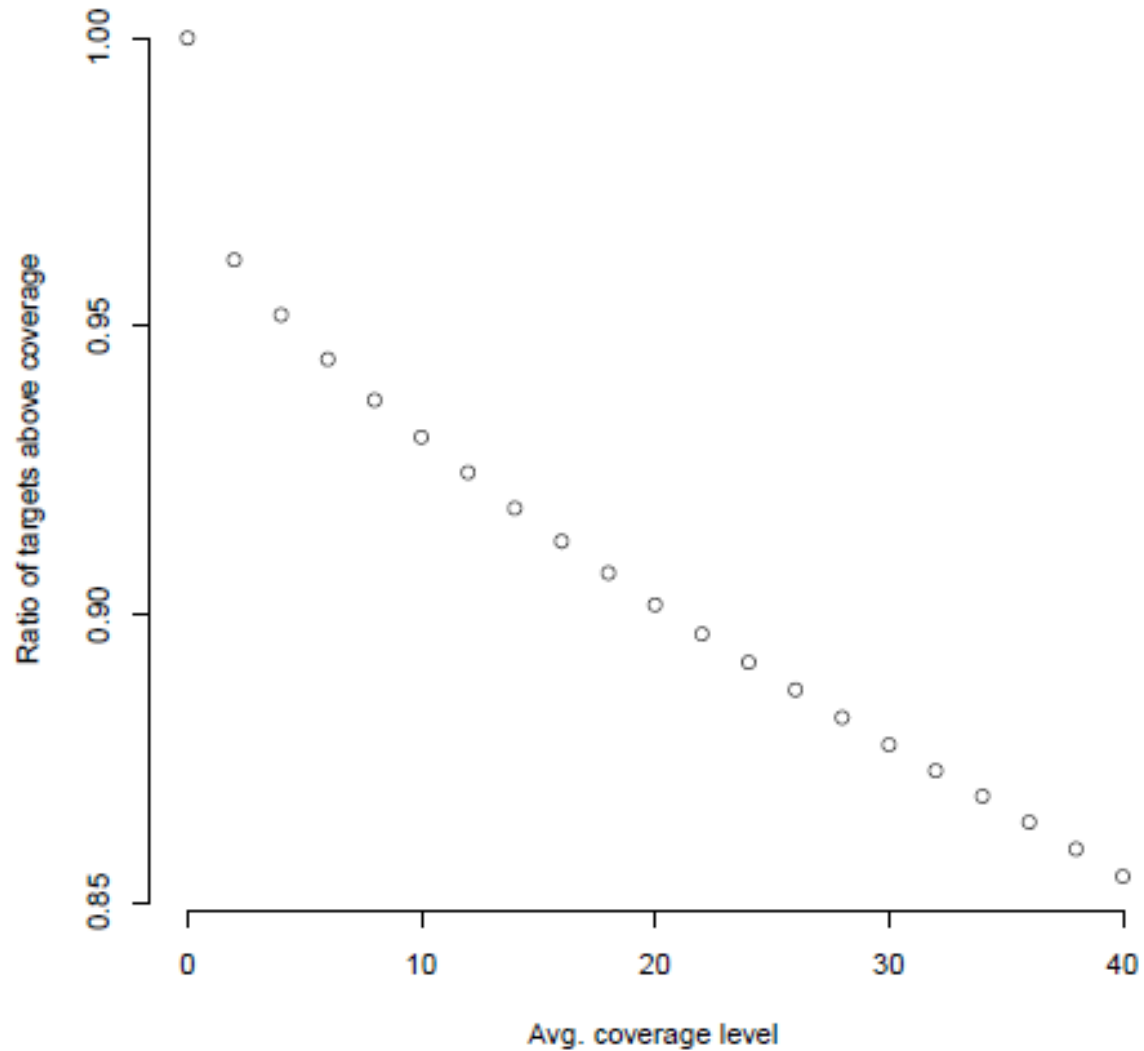
HiSeq Lane, 50MB Kit

s_8_sequence_SureSelect.bed



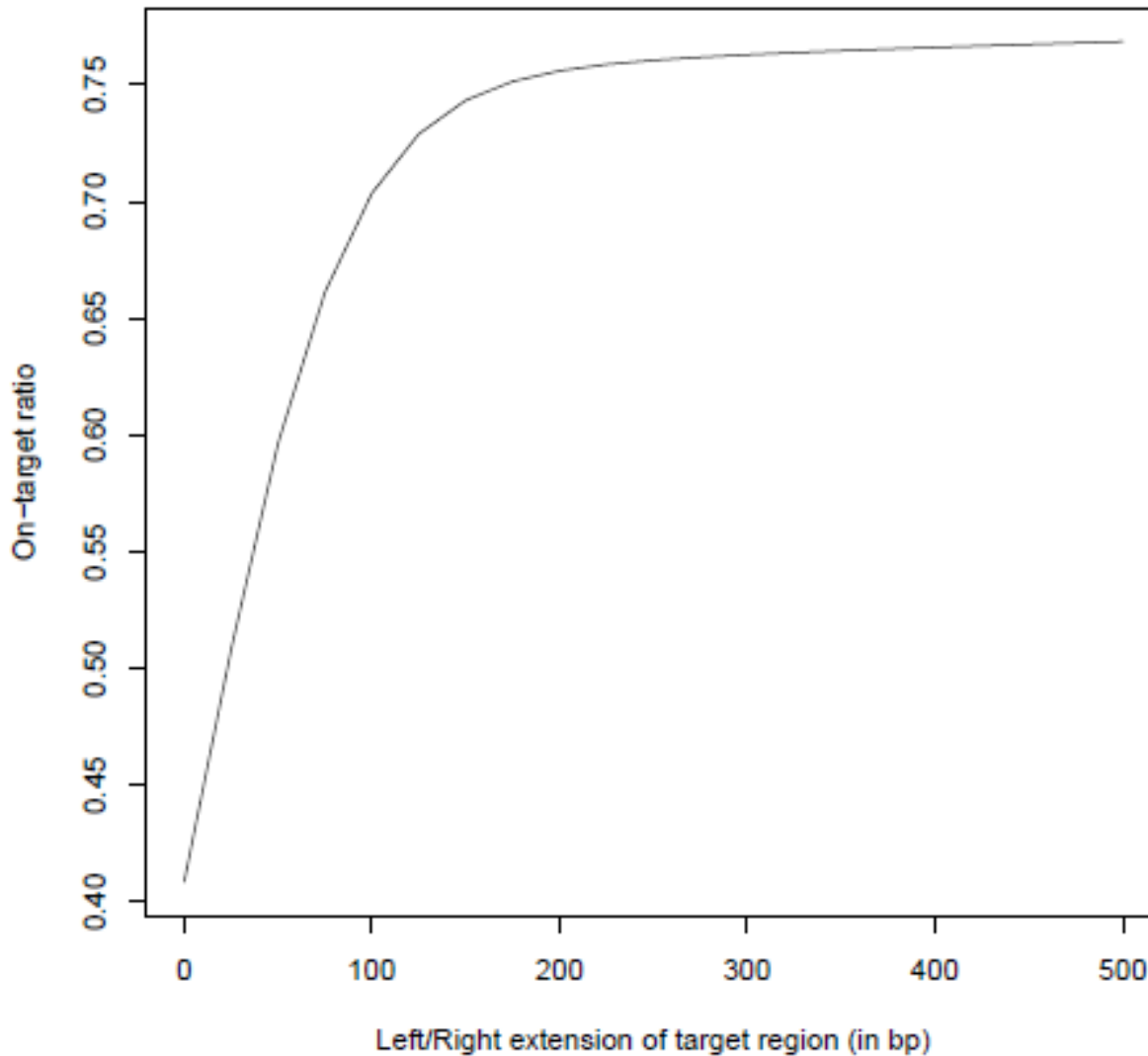
HiSeq Lane, 50MB Kit

s_8_sequence_SureSelect.bed



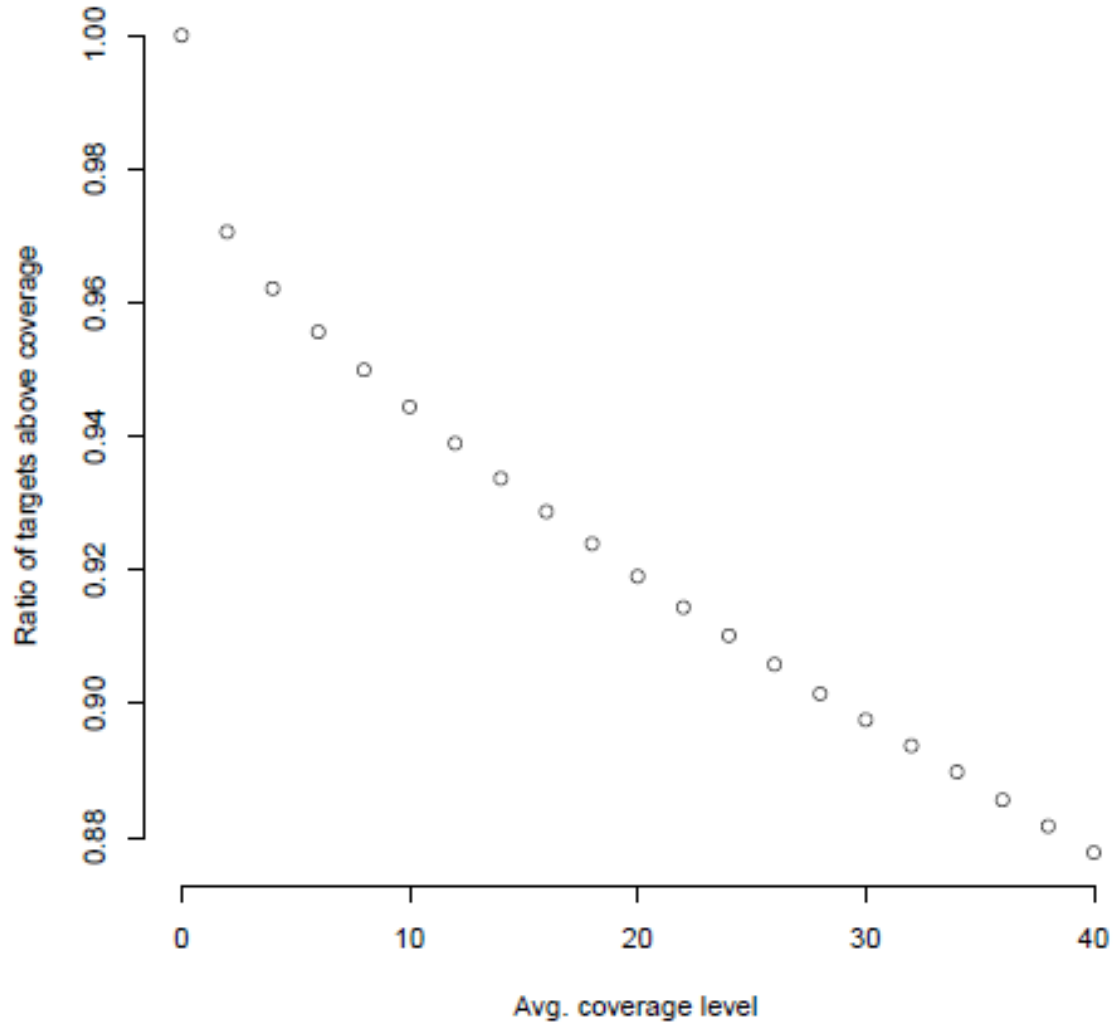
HiSeq Lane, 50MB Kit

s_8_sequence_refseq.bed

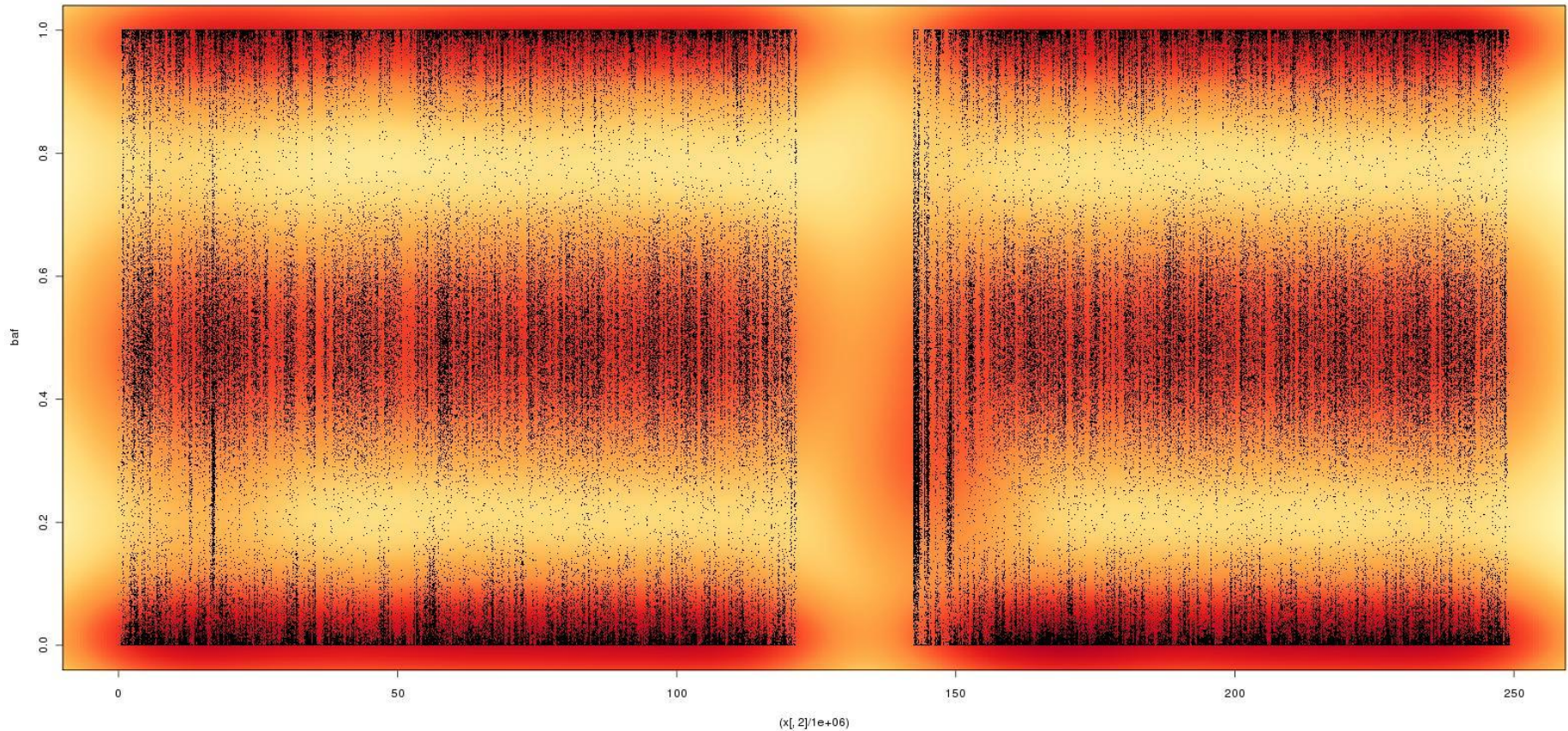


HiSeq Lane, 50MB Kit

s_8_sequence_refseq.bed

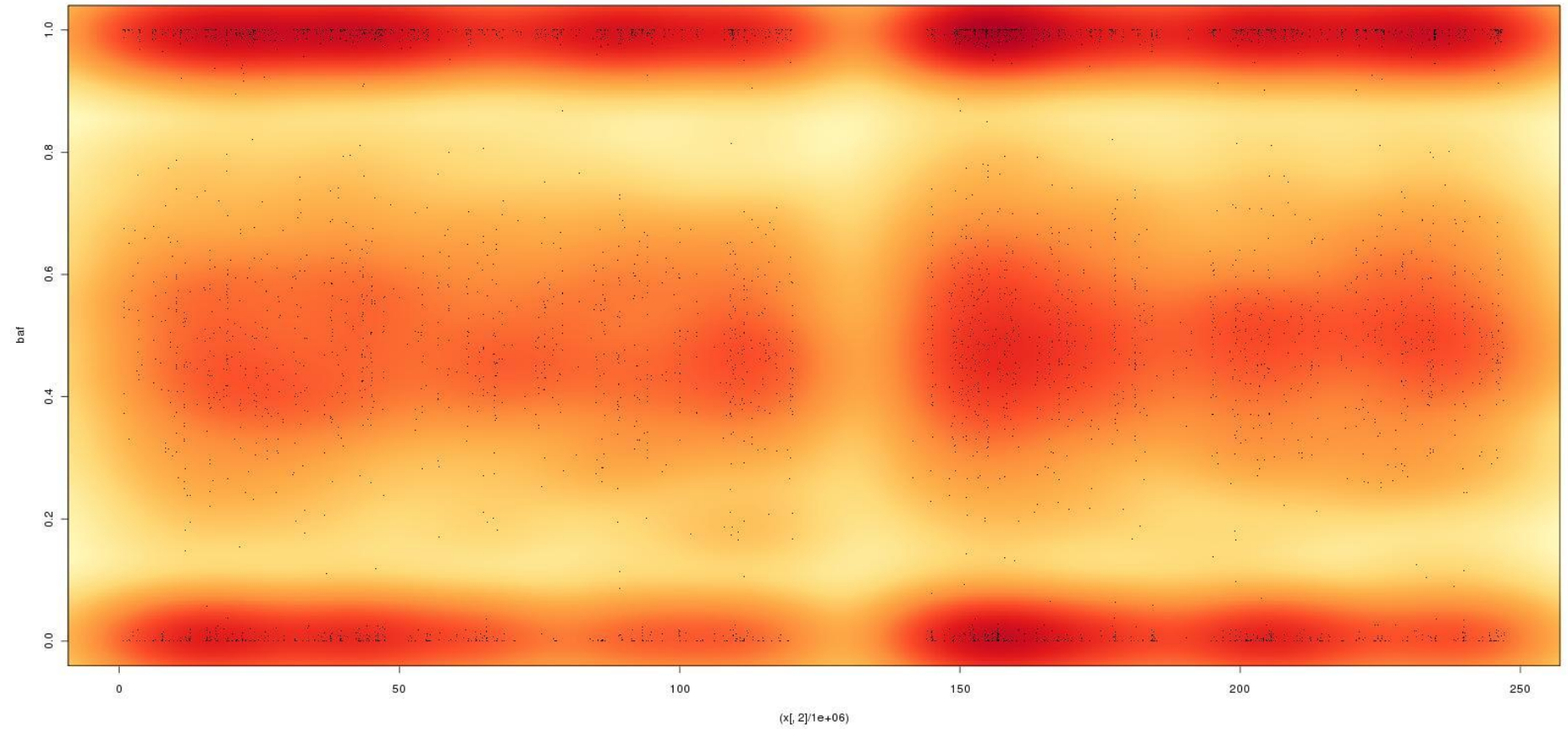


Allelic Balance chr1, whole genome seq.



Allelic Balance

chr1, exon capture



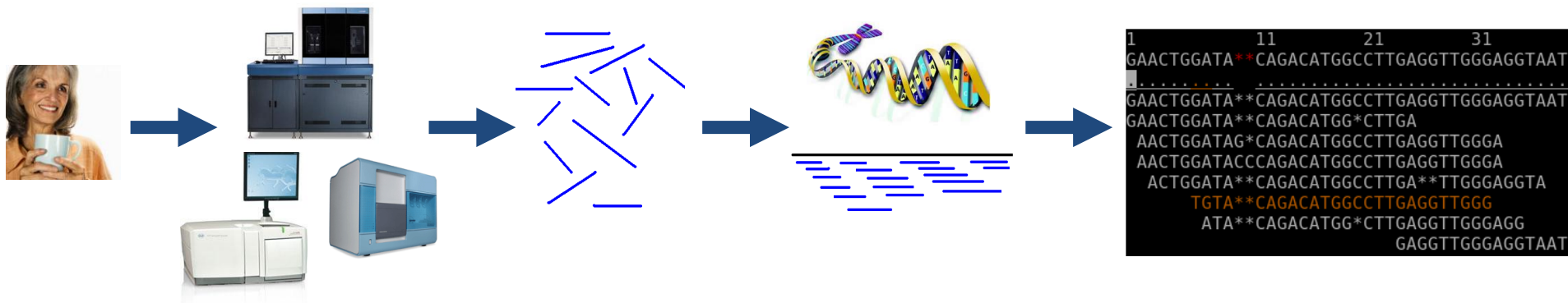
Practical Session

- All shown statistics and summaries are easy to compute using Linux commands and R Statistics.

Welcome to the Practicals!

Material: www.embl.de/~rausch

How to Store Millions of Short-Read Alignments?



Tobias Rausch
June 2011

SAM/BAM

- Generic format for storing large nucleotide sequence alignments
- SAM Tools
 - Sorting alignments
 - Merging alignments
 - Indexing alignments
 - Viewing alignments

SAM record

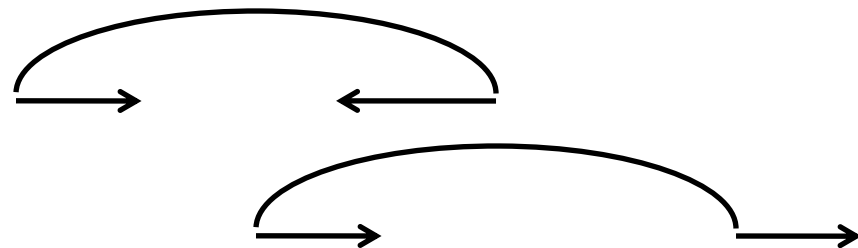
□ Tab-delimited format

- Field 1: Query name
- Field 2: Flag
- Field 3: Reference sequence name
- Field 4: **1-based leftmost** coordinate of the **clipped** sequence
- Field 5: Mapping quality
- Field 6: CIGAR strings
- Field 7: Mate reference sequence name
- Field 8: **1-based leftmost** coordinate of the **clipped** sequence
- Field 9: Insert size (**5' to 5'**)
- Field 10: Query sequence
- Field 11: Sequence qualities

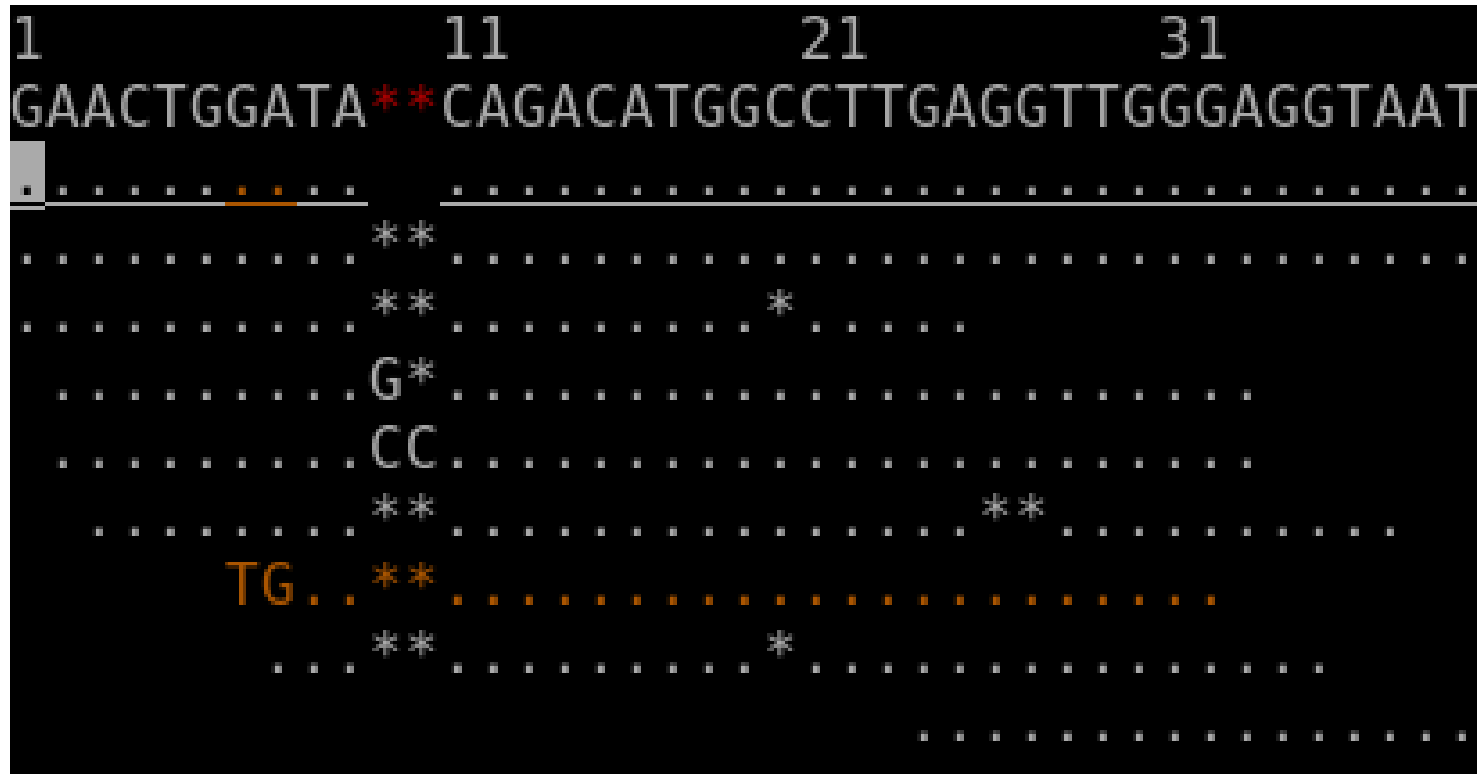
SAM record

□ Tab-delimited format

- Field 1: Query name
- Field 2: Flag
- Field 3: Reference sequence name
- Field 4: **1-based leftmost** coordinate of the **clipped** sequence
- Field 5: Mapping quality
- Field 6: CIGAR strings
- Field 7: Mate reference sequence name
- Field 8: **1-based leftmost** coordinate of the **clipped** sequence
- Field 9: Insert size (**5' to 5'**)
- Field 10: Query sequence
- Field 11: Sequence qualities

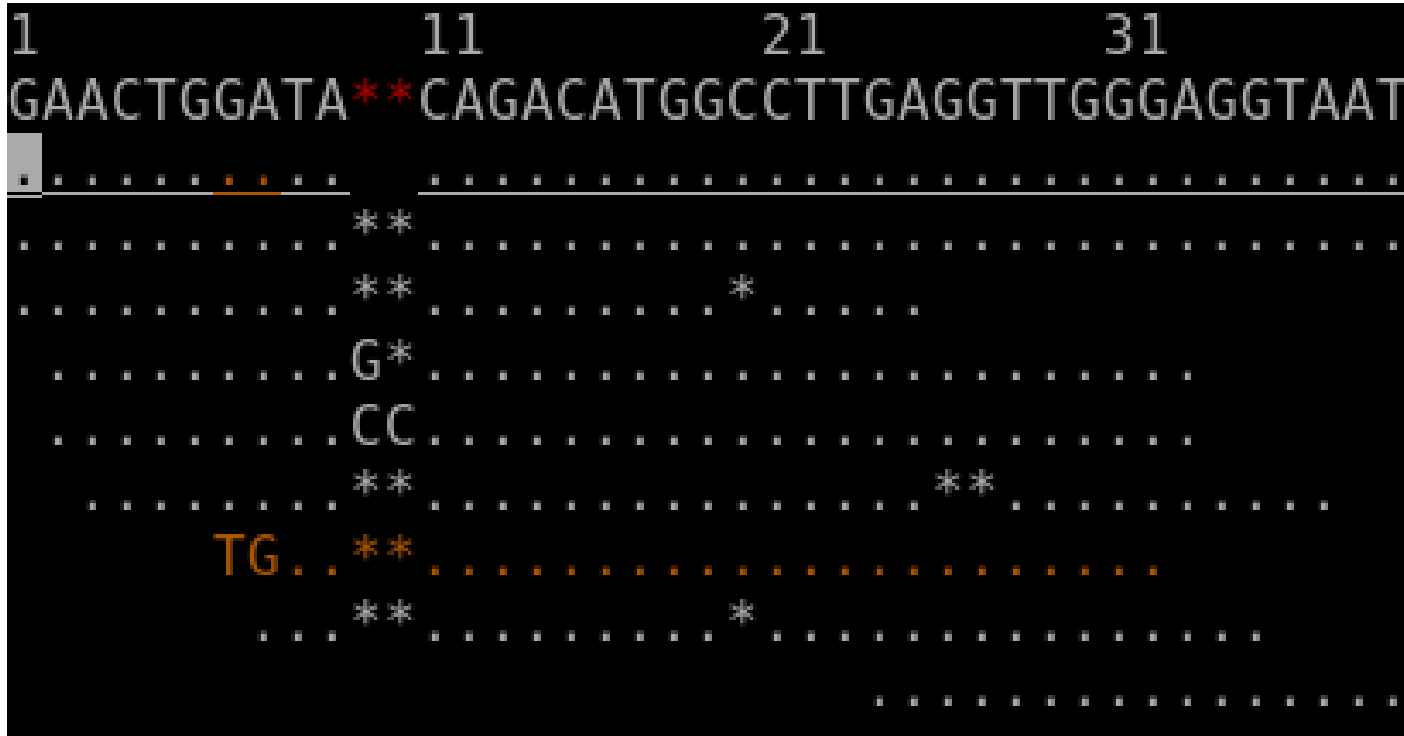


Sam / Bam Format



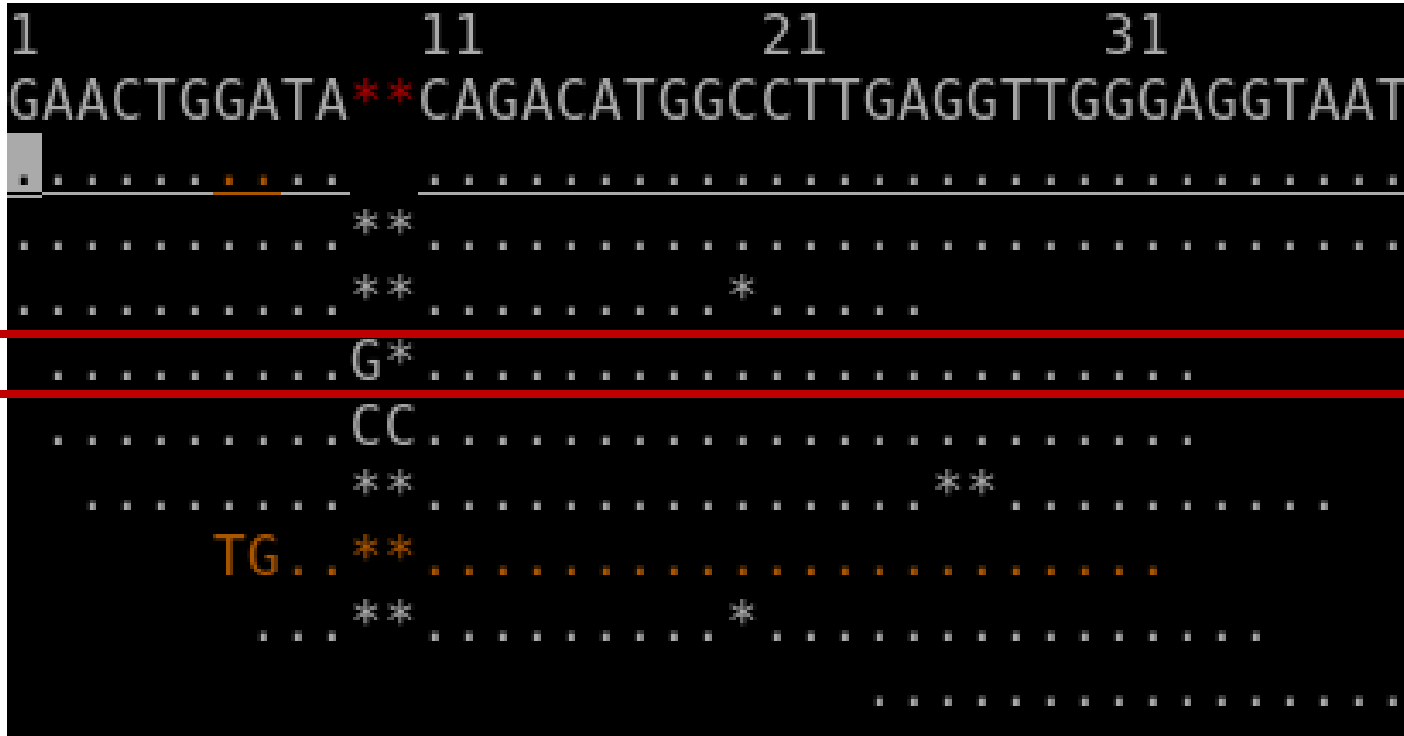
- Sequence characters agreeing with the reference are set to “ . “ or “ , “ for reads aligned to the forward or reverse strand.

Sam / Bam Format



- M: Alignment match or mismatch
- I: Insertion to the reference
- D: Deletion from the reference

Sam / Bam Format



- P: Padding (silent deletion)
- This is not even implemented by BWA
 - Because it would require a *de novo local assembler!*

Sam / Bam Format

- N: Skipped region from the reference
 - For spliced reads:
 - ACATGATA.....GAGCTTTA (Cigar: 8M56N8M)
- Two more CIGAR characters
 - S: Soft clip on the read
 - H: Hard clip on the read

Flags

Bitwise FLAG: $f_{15}f_{14}f_{13}f_{12}f_{11}f_{10}f_9f_8f_7f_6f_5f_4f_3f_2f_1f_0$ with $f_i = \{0,1\}$

f_0 : 0 = Read is not paired in sequencing, 1 = Read is paired in seq.

f_1 : 1 = The read is mapped in a proper pair

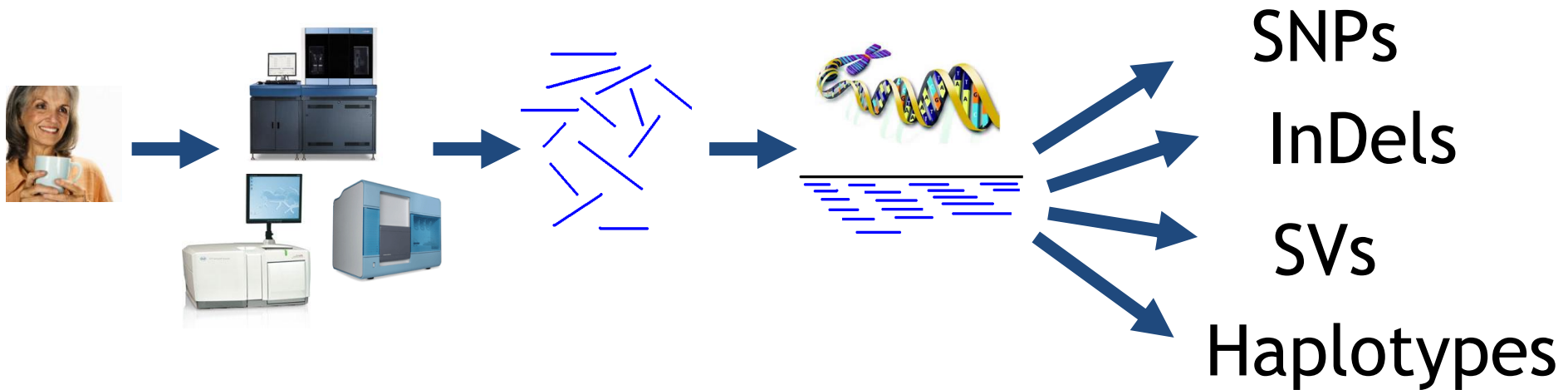
f_2 : 1 = The query sequence itself is unmapped

f_3 : 1 = The mate is unmapped

f_4 : 0 = forward strand, 1 = reverse strand

...

Variant Call Format (VCF)



Tobias Rausch
June 2011

VCFtools

- Methods for working with VCF files
 - Validating vcf files
 - Merging vcf files
 - Comparing vcf files
 - Calculate statistics

VCF version 4.1

- 3 types of information
 - Meta-information lines
 - One header line
 - Data lines, each containing information about a variant in the genome.

Example, SNPs

##fileformat=VCFv4.1

##fileDate=20090805

##FILTER=<ID=q10,Description="Quality below 10">

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00003	...
20	14370	rs605	G	A	29	PASS	DP=14;AF=0.5	GT:GQ:DP:HQ	0 0:48:1:51,51	1/1:43:5:..	
20	17330	.	T	A	3	q10	DP=11;AF=0.017	GT:GQ:DP:HQ	0 0:49:3:58,50	0 1:3:5:65,3	

Example, InDels

##fileformat=VCFv4.1

##fileDate=20090805

##FILTER=<ID=q10,Description="Quality below 10">

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00003
20	123	microsat	GTC	G,GTCT	50	PASS	DP=9	GT:GQ:DP	0/1:35:4	0/2:17:2

Example, SVs

```
##fileformat=VCFv4.1
```

```
#CHROM POS ID REF ALT QUAL FILTER INFO
```

```
2 3216 . N <DEL> 6 PASS IMPRECISE;SVTYPE=DEL;END=3318;SVLEN=102
```