

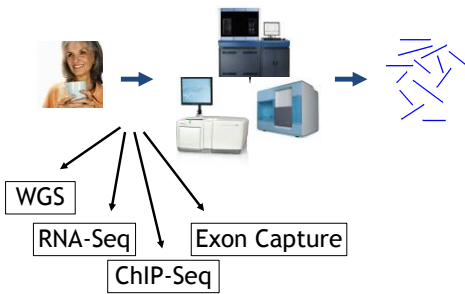
Genome Capture - Data Analysis -

Tobias Rausch
17th June 2010

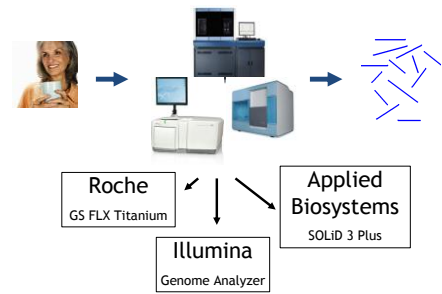
Sequencing



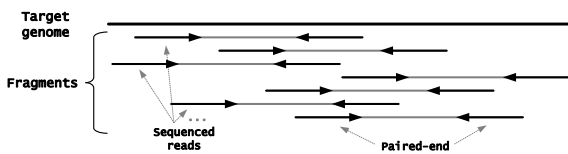
Sequencing



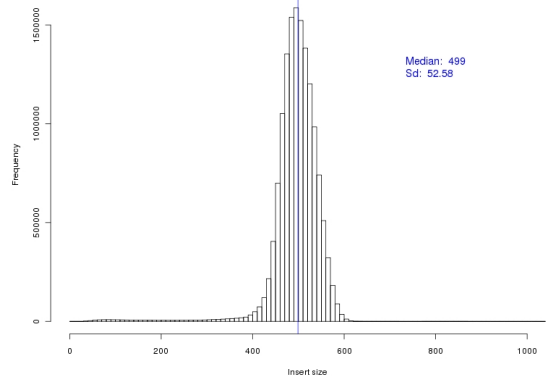
Sequencing



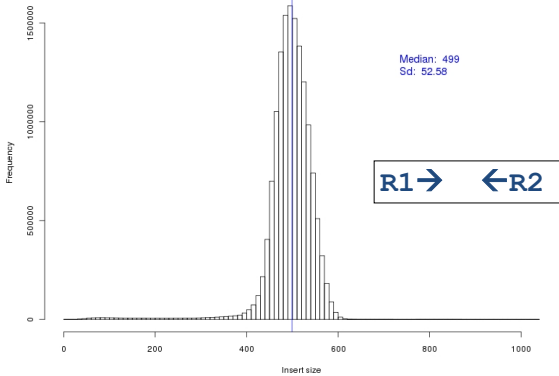
Paired-End Sequencing



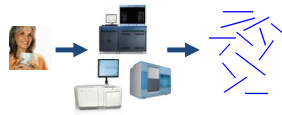
Paired-End Libraries



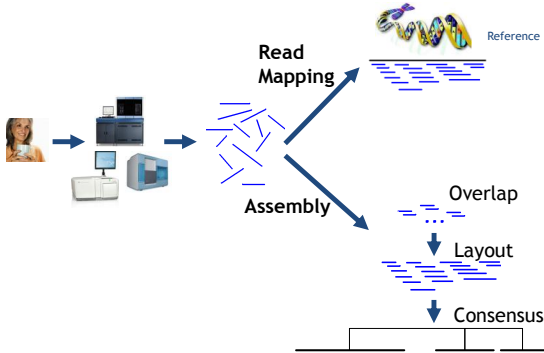
Paired-End Libraries



Data Analysis



Data Analysis



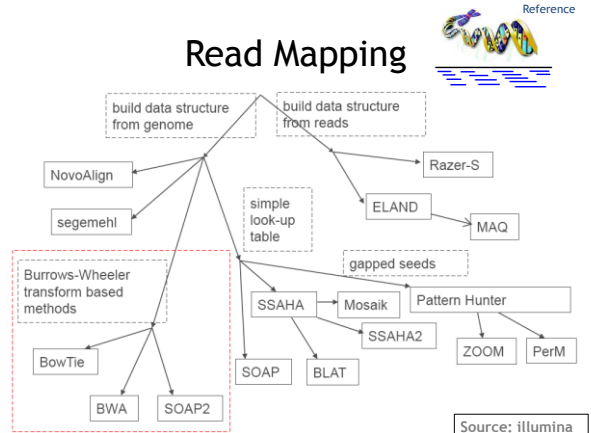
Assembly

- String Graph Assembler
 - Overlap - Layout - Consensus assemblers
 - Examples
 - *Celera Assembler, Arachne, Atlas*
- De-Bruijn Graph Assembler
 - Short-read assemblers
 - Examples:
 - *Velvet, Abyss, SOAPdenovo*
 - Transcriptome assembly: *Oases*

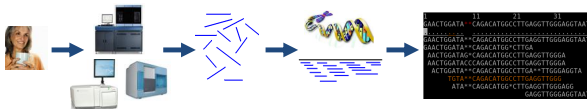
Read Mapping



Read Mapping



How to Store Millions of Short-Read Alignments?



SAM/BAM

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

SAM record

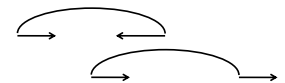
□ Tab-delimited format

- 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

```

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

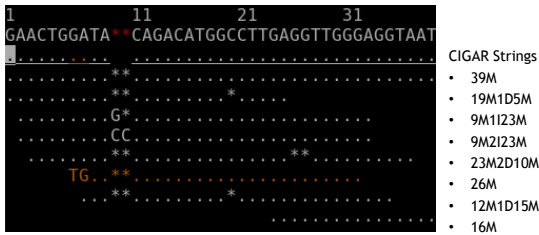
Sam / Bam Format

```

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

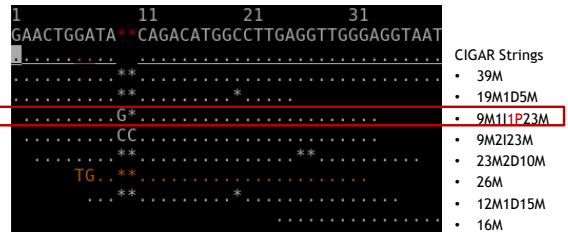
- Sequence characters agreeing with the reference are set to “.” or “,” for reads aligned to the forward or reverse strand.

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

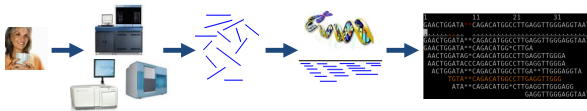
- N: Skipped region from the reference
 - For spliced reads:
 - ACATGATA.....GAGCTTTA (Cigar: 8M56N8M)
- Two more CIGAR characters
 - S: Soft clip on the read
 - H: Hard clip on the read

Flags

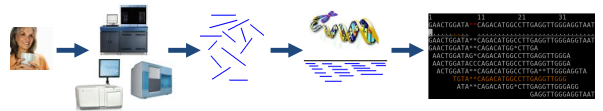
Bitwise FLAG: $f_{15}f_{14}f_{13}f_{12}f_{11}f_{10}f_9f_8f_7f_6f_5f_4f_3f_2f_1f_0$ with $f_i \in \{0,1\}$

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
 ...

Genome Capture Analysis

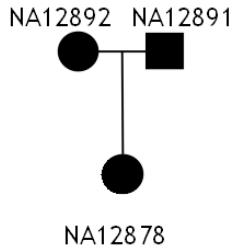


Genome Capture Analysis

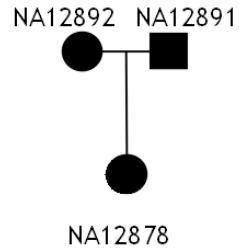


- On-target / Off-target Analysis
- Coverage Analysis
- GC-Content
- SNP Calling
- Relating the Variant Calls to Public Databases

HapMap Trio

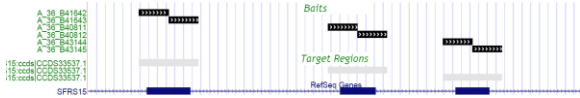


HapMap Trio

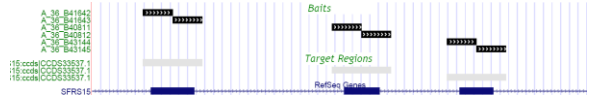


- NA12878 and NA12891 were sequenced

Individual Baits vs. Target Regions

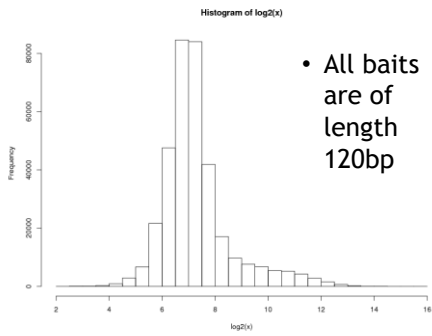


Individual Baits vs. Target Regions



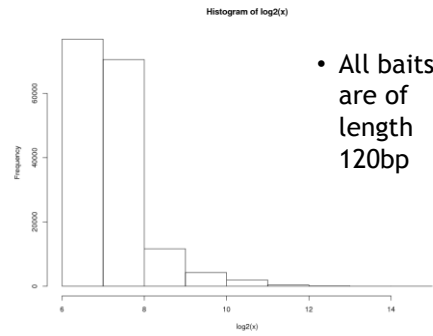
- All bait and target region coordinates are hg18
- Total length of target regions: 37806033 (≈38MB)
- Total length of bait sequences: 38235516 (≈38MB)
- Approximately 1% of the human genome

Exon Length Distribution



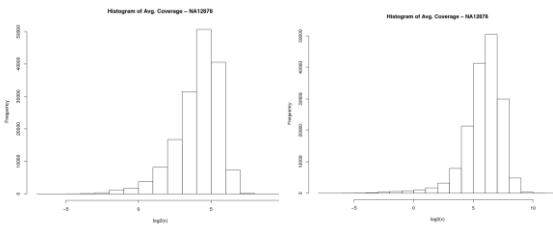
- All baits are of length 120bp

Target Region Length Distribution



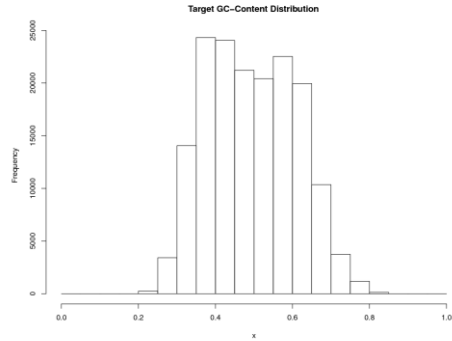
- All baits are of length 120bp

Avg. Coverage for each Target

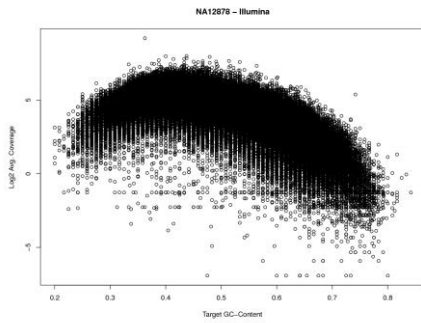


- 3490 Targets without any mapped base
- ← Overlap: 825 →
- 1280 Targets without any mapped base

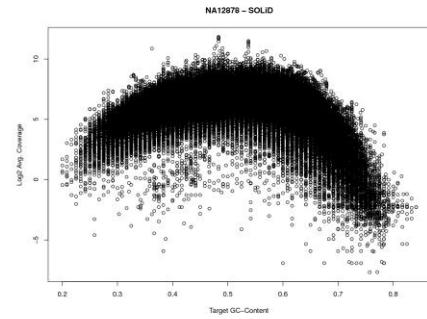
GC-Content Distribution



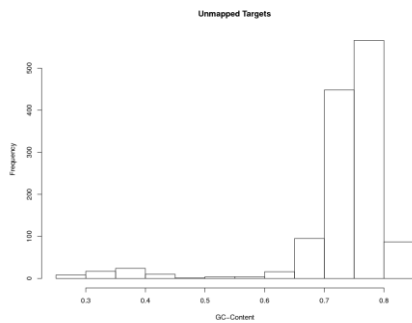
GC-Content Distribution



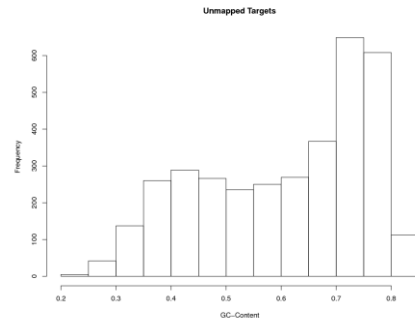
GC-Content Distribution



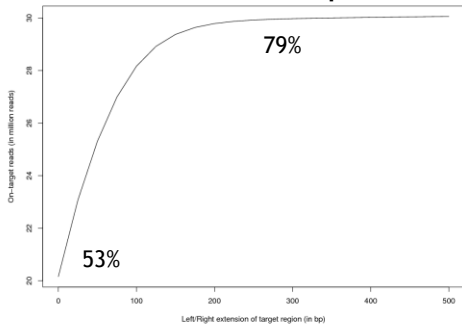
Histogram of GC-Content of Unmapped Targets - SOLiD



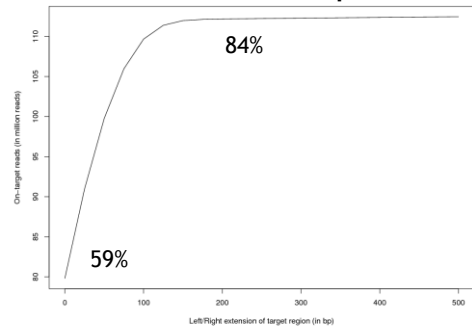
Histogram of GC-Content of Unmapped Targets - Illumina



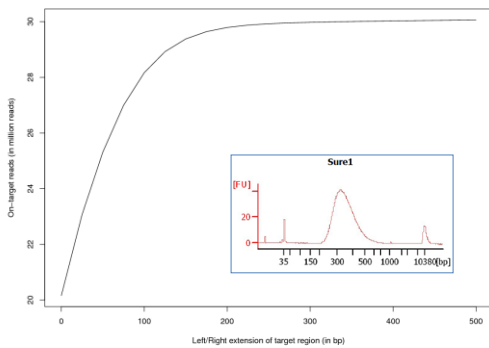
Where did the off-target reads end-up?



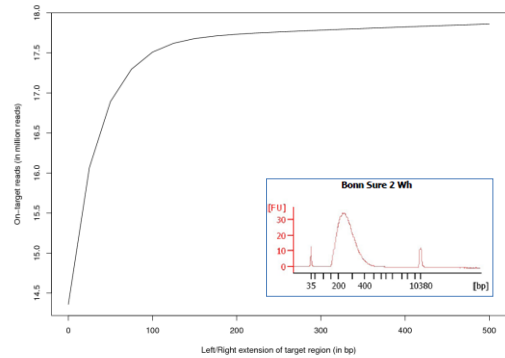
Where did the off-target reads end-up?



Insert Size is very important!



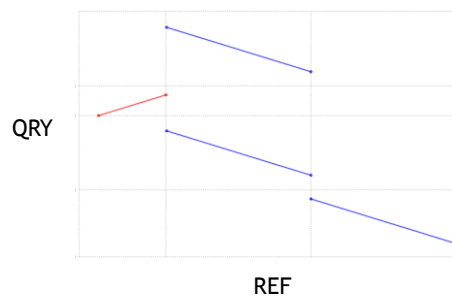
Different Sample: Smaller insert size



Where did the remaining off-target reads end-up?

- Calculate the coverage genome-wide
 - Non-overlapping 100bp windows
 - Select all windows with avg. coverage ≥ 10
 - Merge subsequent windows of coverage ≥ 10
 - NA12878 : 126681 regions
 - NA12891 : 124057 regions
 - Deduce all regions overlapping with one of the target regions
 - Compare the remaining high-coverage regions with missed target regions using MUMmer

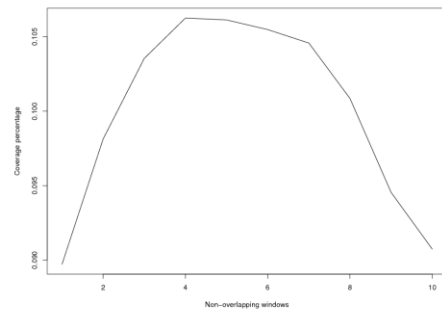
QRY: High-coverage genome regions Ref: Missed target regions



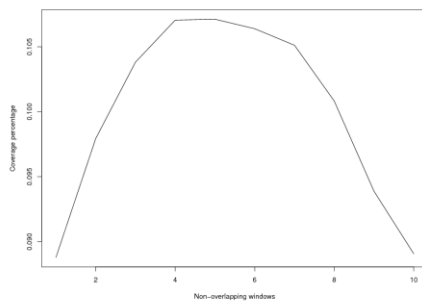
Uniform Coverage across Targets?

- 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

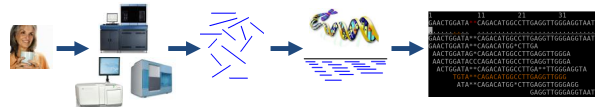
Uniform Coverage across Targets? NA12878 - Illumina



Uniform Coverage across Targets? NA12891 - Illumina

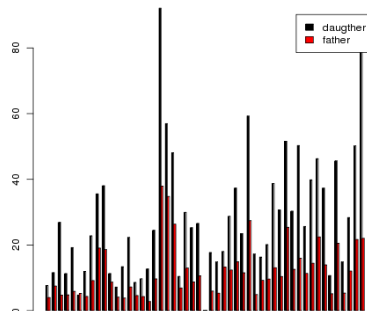


Genome Capture Analysis



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

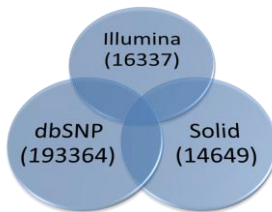
50 random targets on chrX



Fraction of Reads on chrX

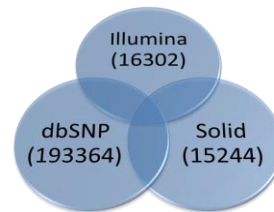
- Illumina data
 - NA12878: 0.02475
 - NA12891: 0.01329
- SOLiD data
 - NA12878: 0.02995
 - NA12891: 0.01162

dbSNP - NA12878



- About 3 times as much data for SOLiD, however:
 - Illumina & dbSNP: 15379 (94%)
 - Solid & dbSNP: 12513 (96%)
 - Solid & Illumina: 12299 & dbSNP: 12120 (99%)

dbSNP - NA12891



- About 3 times as much data for SOLiD, however:
 - Illumina & dbSNP: 15326 (94%)
 - Solid & dbSNP: 14297 (94%)
 - Solid & Illumina: 12956 & dbSNP: 12734 (98%)

Summary

- Agilent's SureSelect Target Enrichment System seems to work nicely
- On-target ratio is good
- Results are re-producible even across different sequencing platforms
- Short insert-sizes are beneficial
- Paired-end vs. single-end? I don't know.
 - Redundancy is a lot easier to estimate for paired-end data.

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