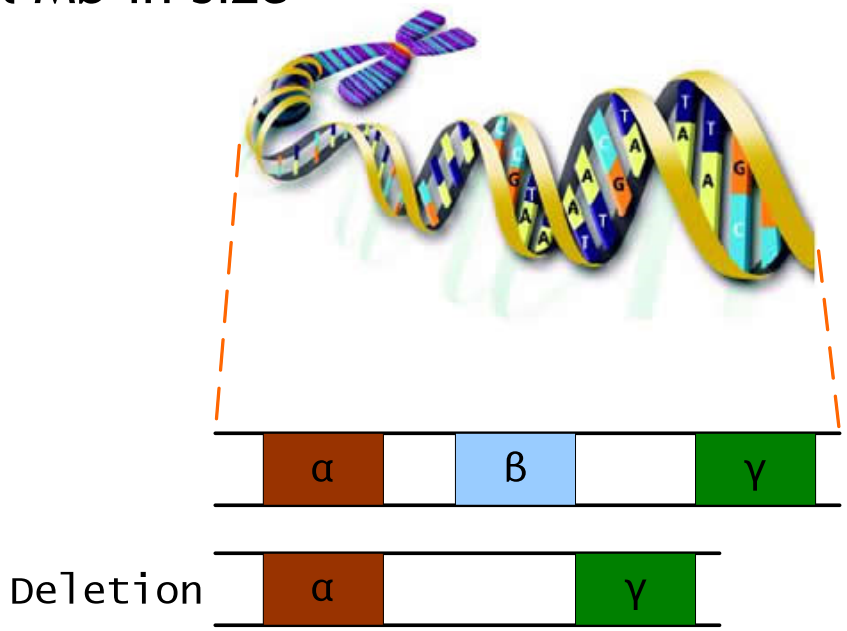
Discovery of Genomic Structural Variations with Next-Generation Sequencing Data

Advanced Topics in Computational Genomics

Slides from Marcel H. Schulz,
Tobias Rausch (EMBL),
and Kai Ye (Leiden University)

Genomic Rearrangements/ Structural Variations (SVs)

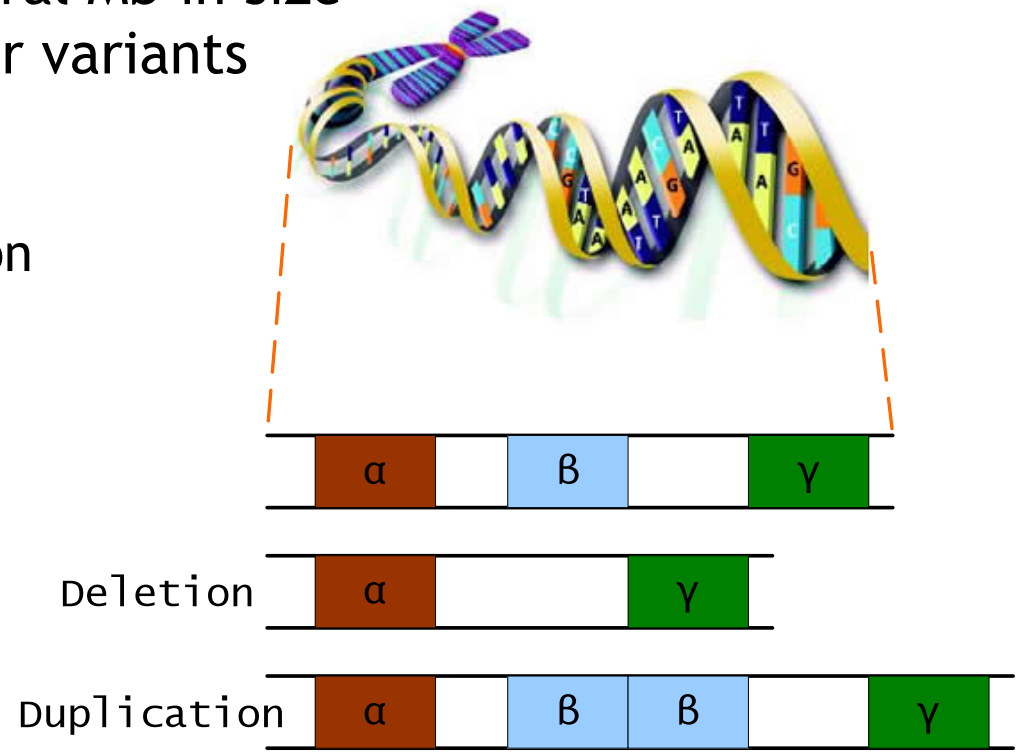
- 1 Kb to several Mb in size



courtesy of Tobias Rausch (EMBL)

Genomic Rearrangements/ Structural Variations (SVs)

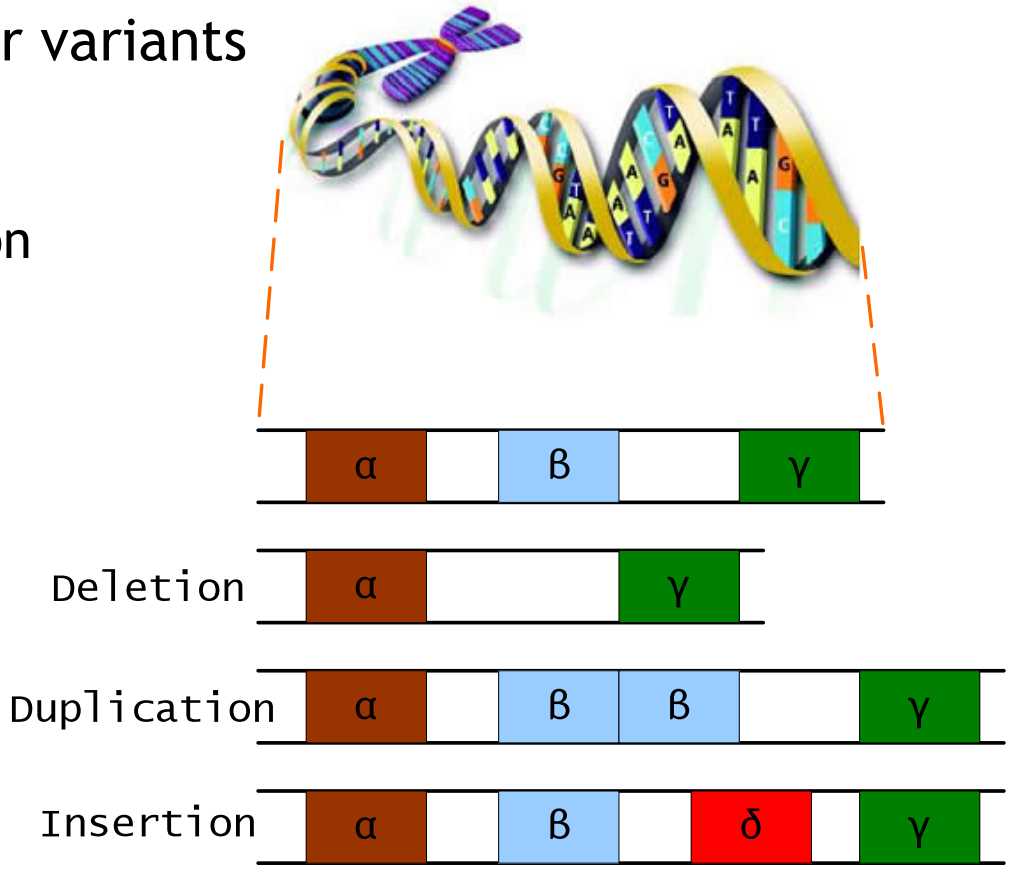
- 1 Kb to several Mb in size
- Copy number variants (CNVs)
  - Deletion
  - Duplication



courtesy of Tobias Rausch (EMBL)

Genomic Rearrangements/ Structural Variations (SVs)

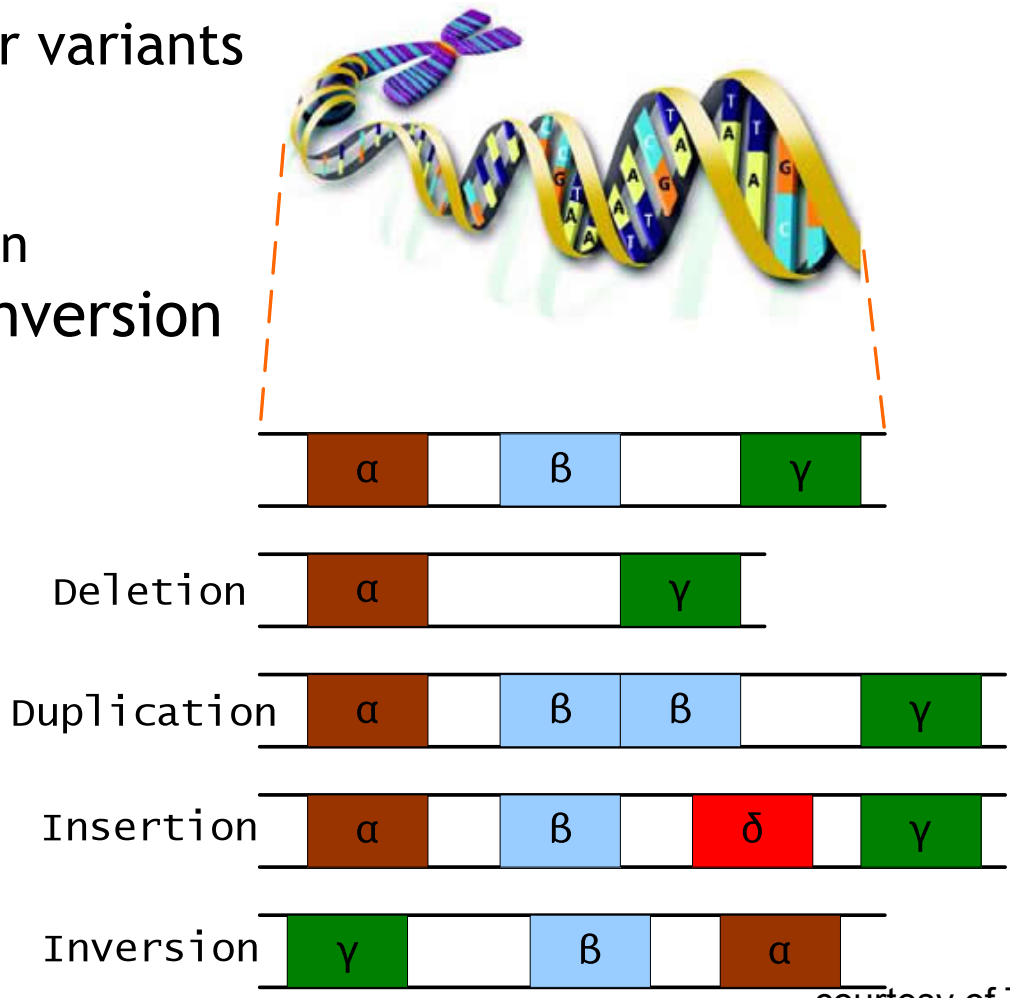
- 1 Kb to several Mb in size
- Copy number variants (CNVs)
  - Deletion
  - Duplication
- Insertion



courtesy of Tobias Rausch (EMBL)

Genomic Rearrangements/ Structural Variations (SVs)

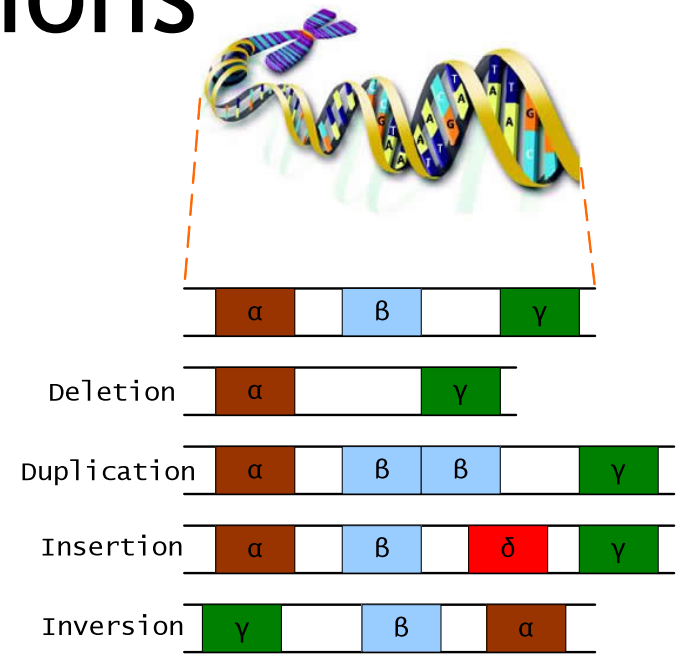
- 1 Kb to several Mb in size
- Copy number variants (CNVs)
  - Deletion
  - Duplication
- Insertion, Inversion



courtesy of Tobias Rausch (EMBL)

Genomic Rearrangements/ Structural Variations

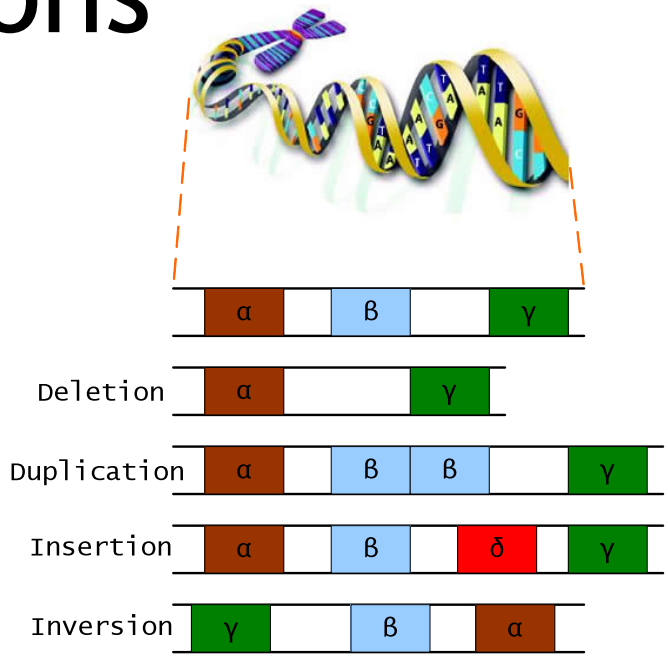
- 1 Kb to several Mb in size
- Copy number variants (CNVs)
  - Deletion
  - Duplication
- Insertion, Inversion, Translocation



courtesy of Tobias Rausch (EMBL)

Genomic Rearrangements/ Structural Variations

- 1 Kb to several Mb in size
- Copy number variants
  - Deletion
  - Duplication
- Insertion, Inversion, Translocation
- More abundant than SNPs



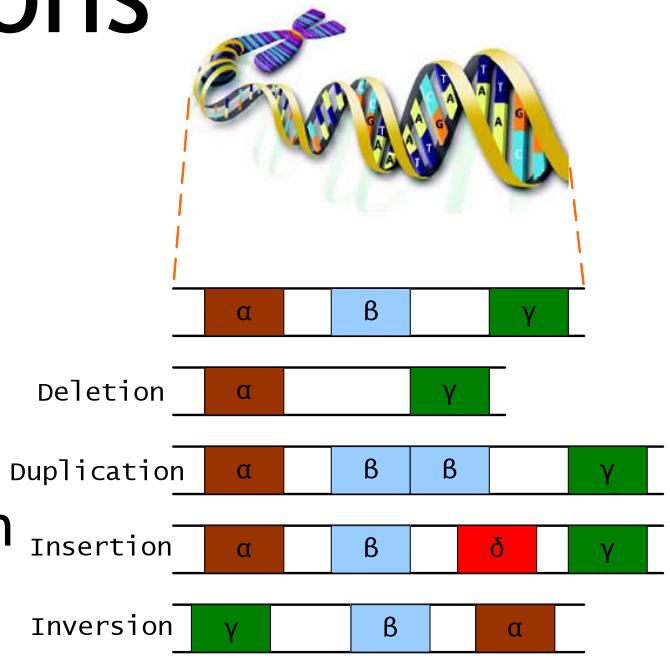
...ACGA**T**ACG...
...ACGA**G**ACG...

	SNPs	CNVs
Base pairs	2.5 Mb	4 Mb
% genome	0.08%	0.12%

courtesy of Tobias Rausch (EMBL)

Genomic Rearrangements/ Structural Variations

- 1 Kb to several Mb in size
- Copy number variants
  - Deletion
  - Duplication
- Insertion, Inversion, Translocation
- More abundant than SNPs
- Either neutral or non-neutral in function
- Non-neutral mechanisms
  - Disrupting genes
  - Creating fusion genes
  - Copy number changes of dosage-sensitive genes



courtesy of Tobias Rausch (EMBL)

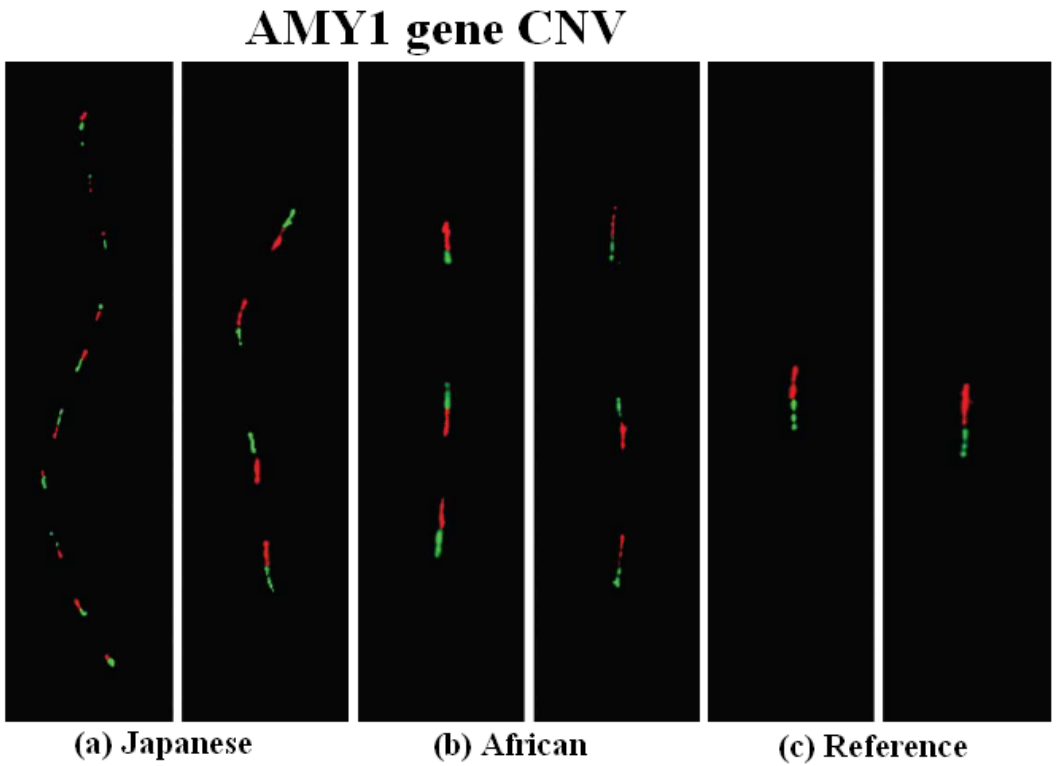
Why Structural Variation Discovery

- Finding disease causal genes
- Trace evolutionary genome history
- Analyze the mechanisms of SVs occurrence
- Understand Repetitive Element spreading (LINEs, ALUs, etc.)

Technologies to Discover Structural Variations

Technologies

- Fluorescent in situ hybridization (FISH)
  - Fluorescent probes (≈100kb) detect and localize the presence or absence of specific DNA sequence
  - Probe should be large enough for a specific hybridization

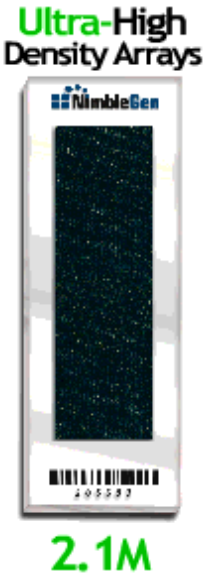


❑Perry et al. (2007)

courtesy of Tobias Rausch (EMBL)

Technologies

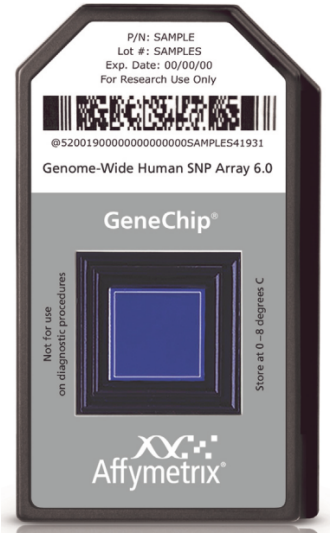
- Fluorescent in situ hybridization (FISH)
- Comparative Genomic Hybridization (CGH)
  - Test vs. reference sample
  - 2.1 million probes
  - Different types
    - Whole-Genome Tiling Arrays
    - Whole-Genome Exon-Focused Arrays
    - CNV Arrays



courtesy of Tobias Rausch (EMBL)

Technologies

- Fluorescent in situ hybridization (FISH)
- Comparative Genomic Hybridization (CGH)
- Genome-Wide Human SNP Array 6.0
  - 1.8 million genetic markers
    - 906,600 SNPs
    - 946,000 probes for CNVs



courtesy of Tobias Rausch (EMBL)

Technologies

- Fluorescent in situ hybridization (FISH)
- Comparative Genomic Hybridization (CGH)
- Genome-Wide Human SNP Array 6.0
- Human 1M-Duo DNA Analysis BeadChip
  - 1.2 million genetic markers
    - Markers for SNPs and CNV regions
  - Targeted studies
    - 60,800 additional custom SNPs
    - 60,000 custom CNV-targets

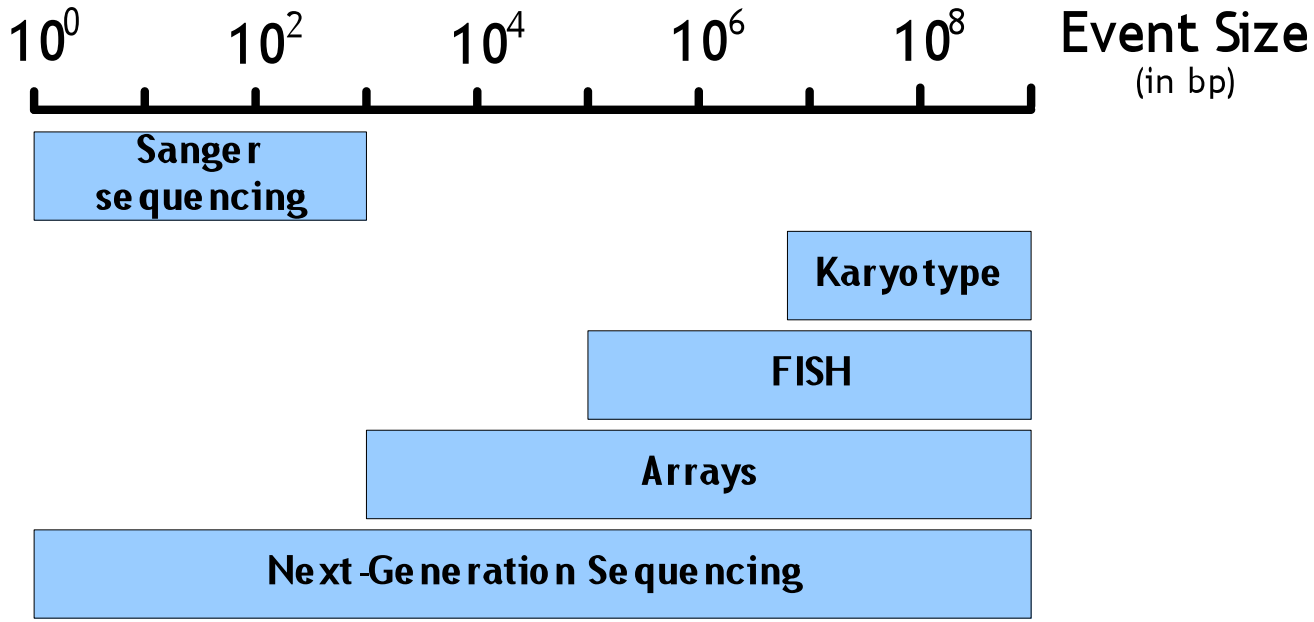
courtesy of Tobias Rausch (EMBL)

Technologies

- Fluorescent in situ hybridization (FISH)
- Comparative Genomic Hybridization (CGH)
- Genome-Wide Human SNP Array 6.0
- Human 1M-Duo DNA Analysis BeadChip
- Next-Generation Sequencing (NGS)
  - Whole-genome sequencing
  - Targeted, e.g. RNA-Seq

courtesy of Tobias Rausch (EMBL)

Focus on NGS

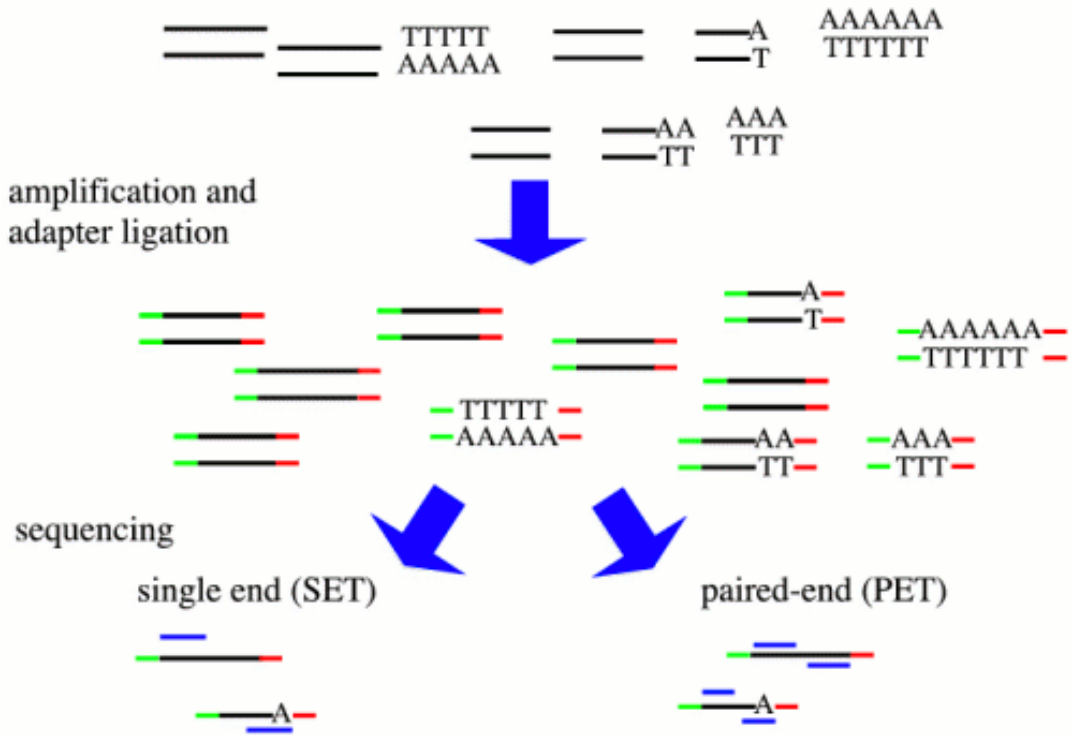


- Limitations of Arrays
  - Lower resolution for genomic rearrangements
  - Balanced events (e.g., inversions) cannot be detected using signal intensity differences
  - No breakpoint information

courtesy of Tobias Rausch (EMBL)

Paired-end data

- Two protocols for paired-end data
  - mate-pair sequencing by circularization (traditional Sanger sequencing)
  - paired-end NGS



overview protocol

Paired-end data

– paired-end NGS (insert distribution known due to fragment size selection)