# PALMapper: Fast, Accurate and Variation-aware RNA-seq Alignments

André Kahles, Géraldine Jean, David Kuo, Vipin T. Sreedharan, Gunnar Rätsch

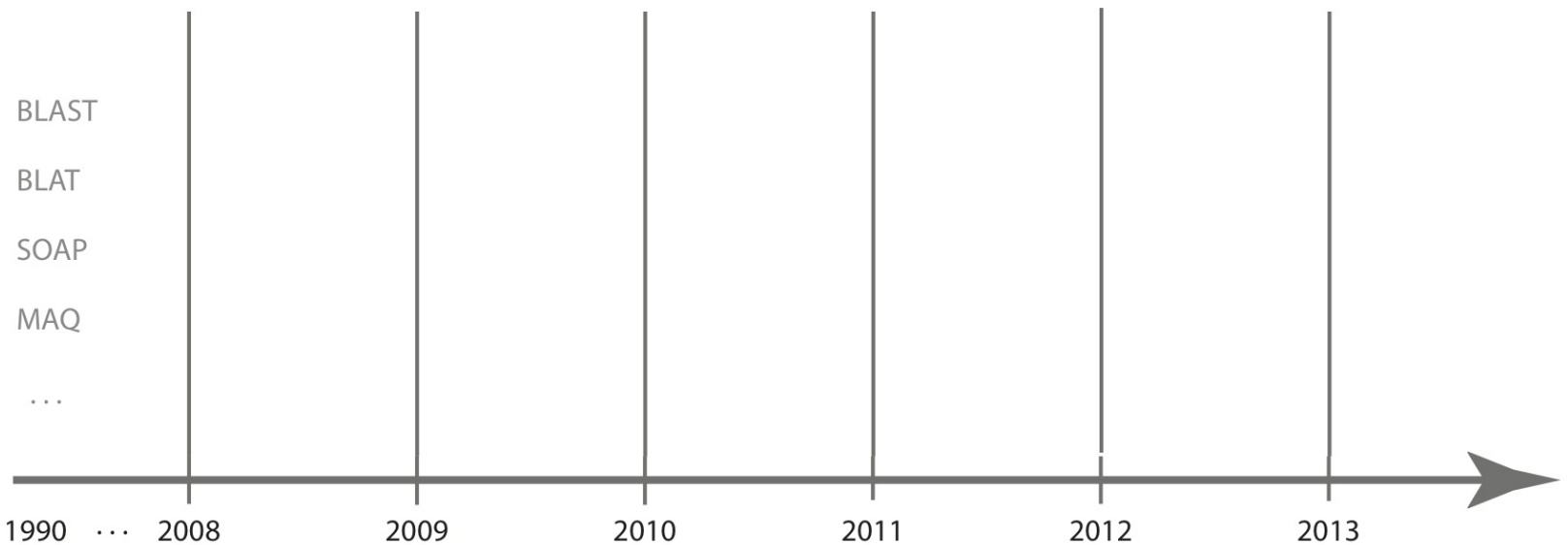

Memorial Sloan-Kettering Cancer Center

New York City, USA

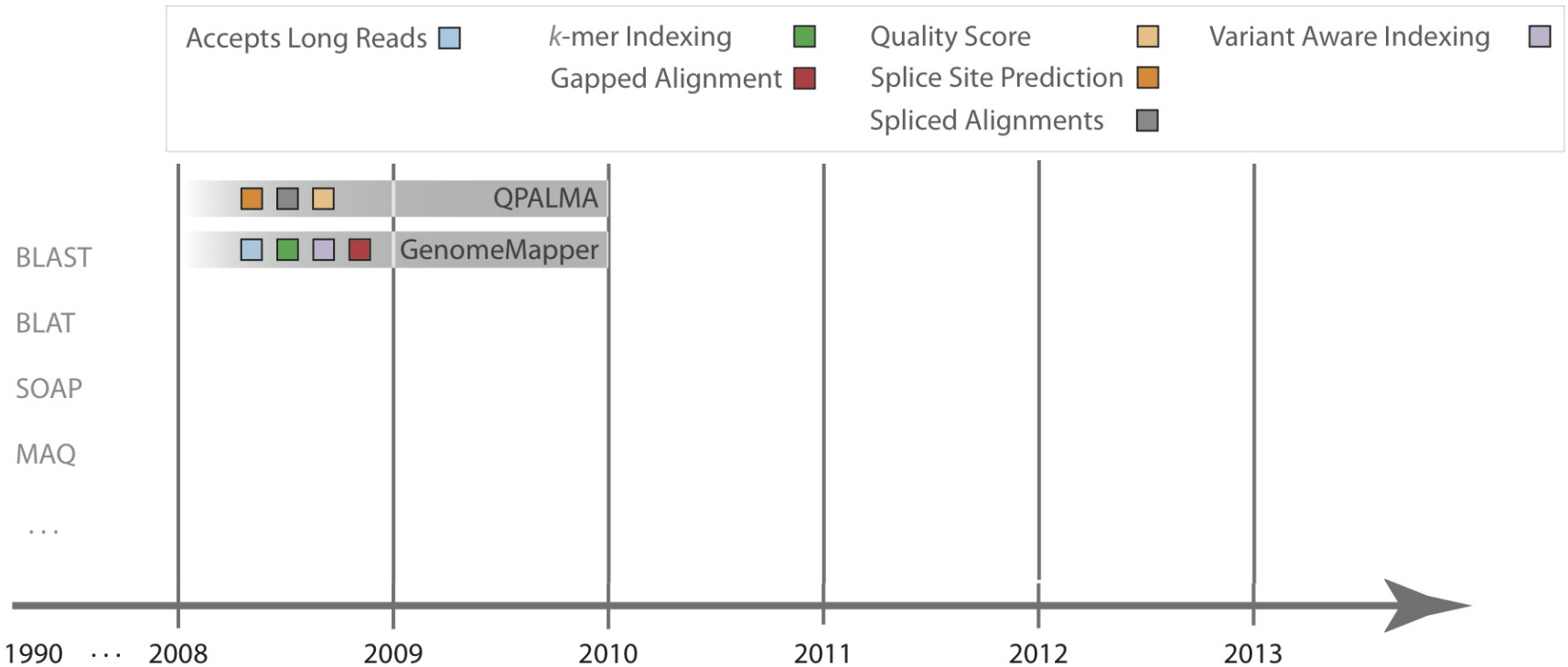
HiTSeq, Berlin, July 20, 2013

# Why Another Mapper?

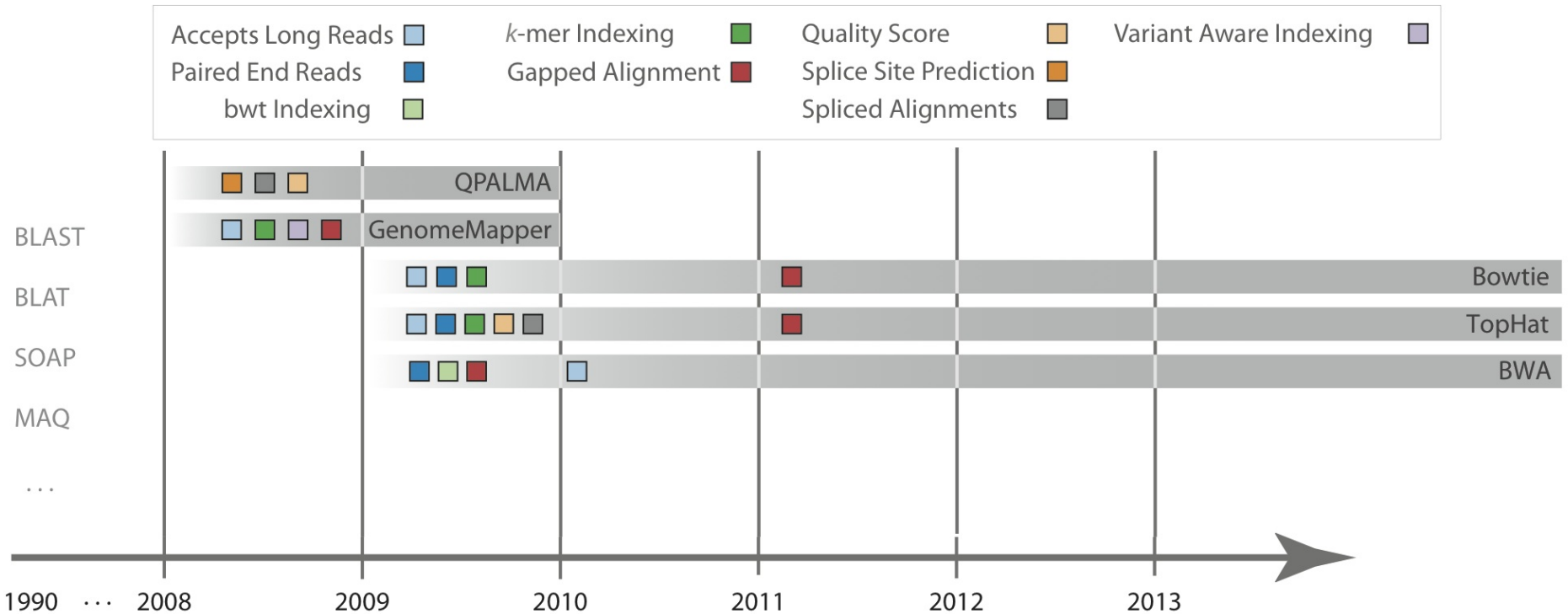
# Why Another Mapper?



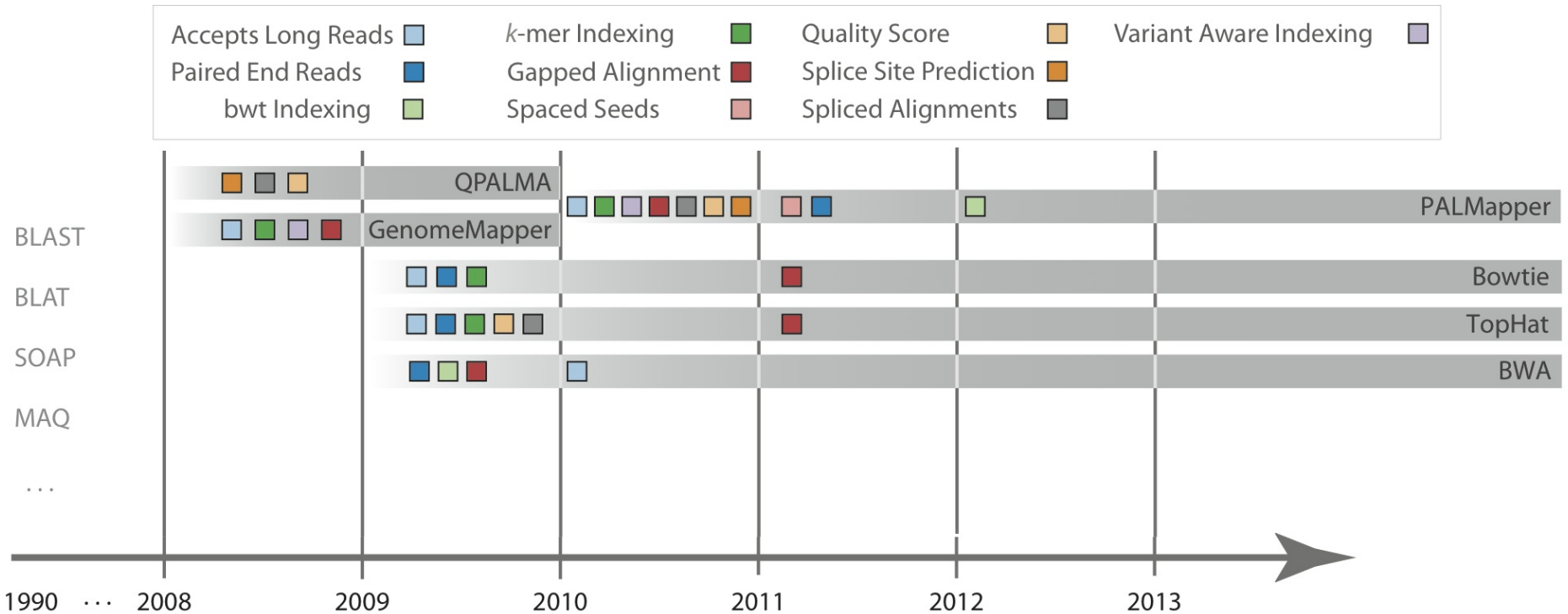
# Why Another Mapper?



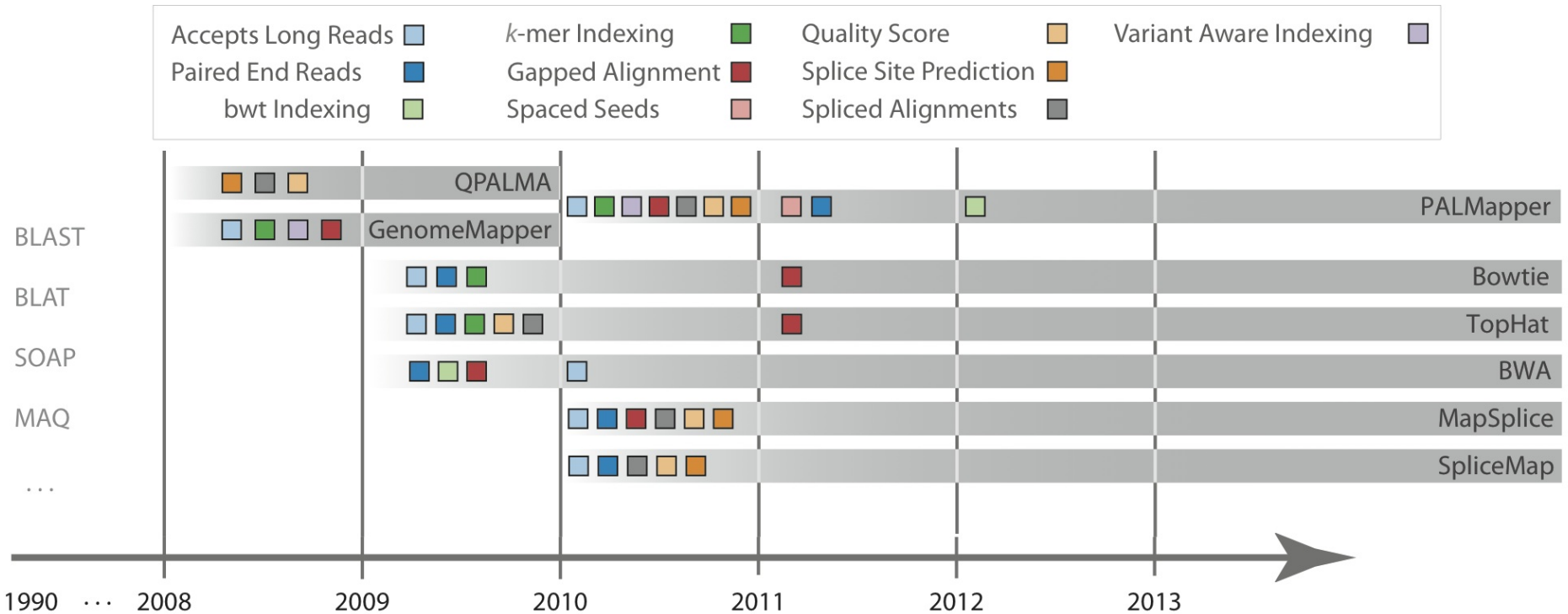
# Why Another Mapper?



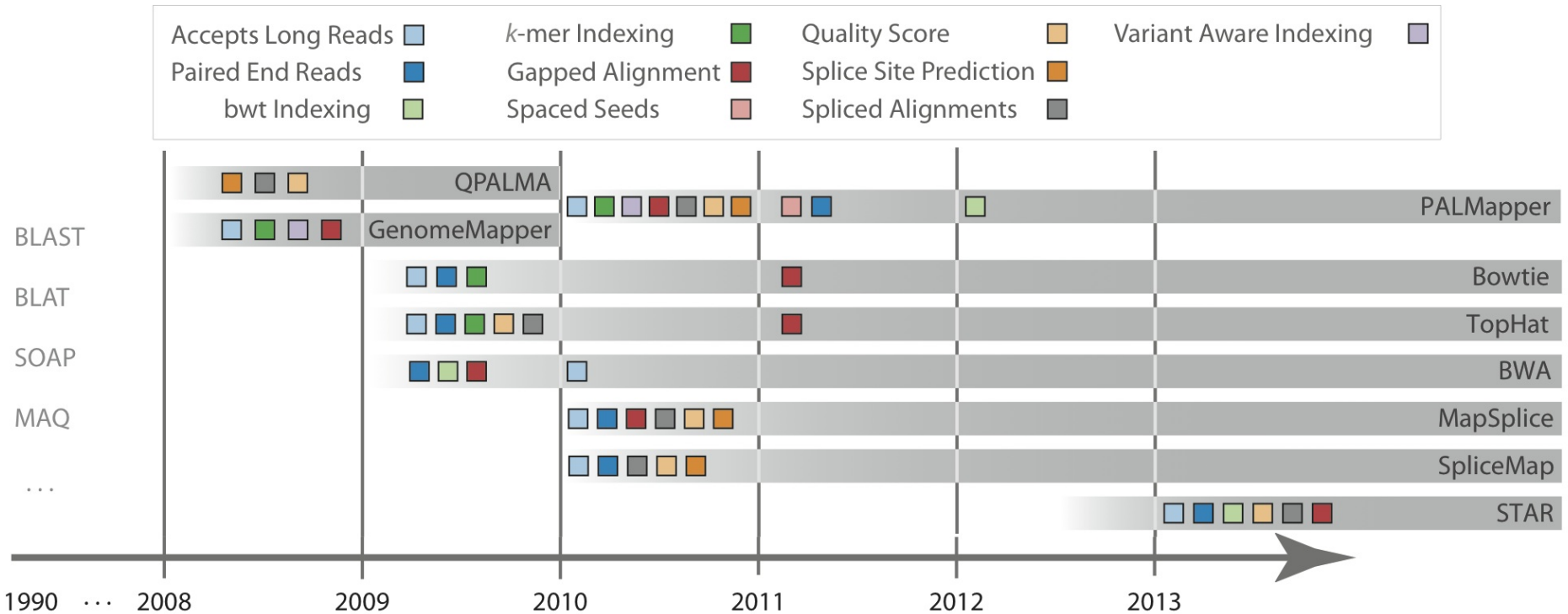
# Why Another Mapper?



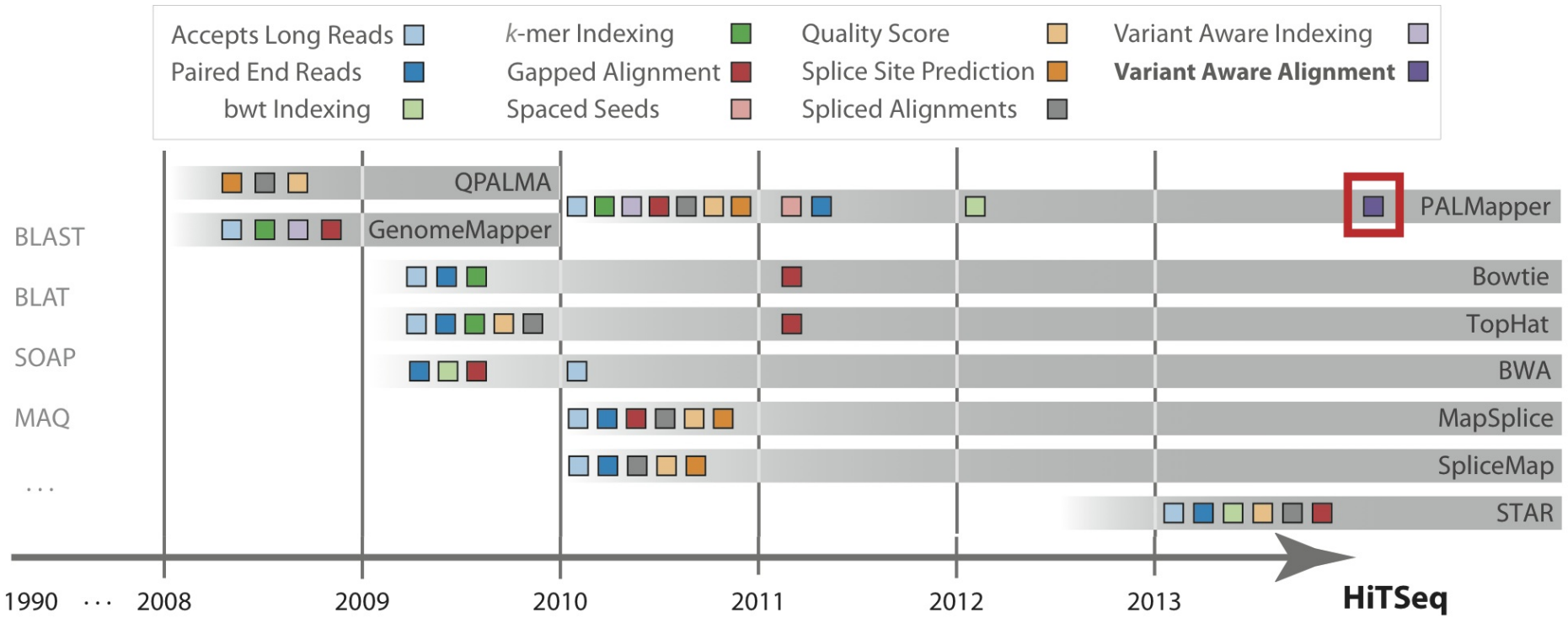
# Why Another Mapper?



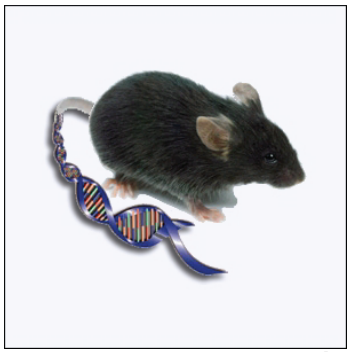
# Why Another Mapper?



# Why Another Mapper?



# Applications for Variation-aware Alignments



source: sanger.ac.uk

Mouse
Genome Project


source: genome.sfu.ca/mmp/

Arabidopsis
1001 Genomes

source: 1001genomes.org

Panzea

source: alpha.maizegdb.org

1000 Genome Project
source: nature.com

| Resource Name | Organism | Strains/Individuals | SNPs/SNVs | InDels | Reference |
|---|---|---|---|---|---|
| Mouse Genome Project | *M. musculus* | 17 strains | 129,260,574 | 21,683,297 | Keane (2011) |
| 19 Genomes Project | *A. thaliana* | 19 strains | 3,070,000 | 1,200,000 | Gan (2011) |
| Panzea | *T. mays* | 103 strains | 55,061,920 | 3,200,000 | Chia (2012) |
| Million Mutation Project | *C. elegans* | 2,047 strains | 800,000 | 220,000 | Thompson (2013) |
| 1000 Genomes Project | *H. sapians* | 1,092 individuals | 36,700,000 | 1,380,000 | The 1000 Genomes Project Consortium (2012) |

# RNA-seq Read Alignments with PALMapper

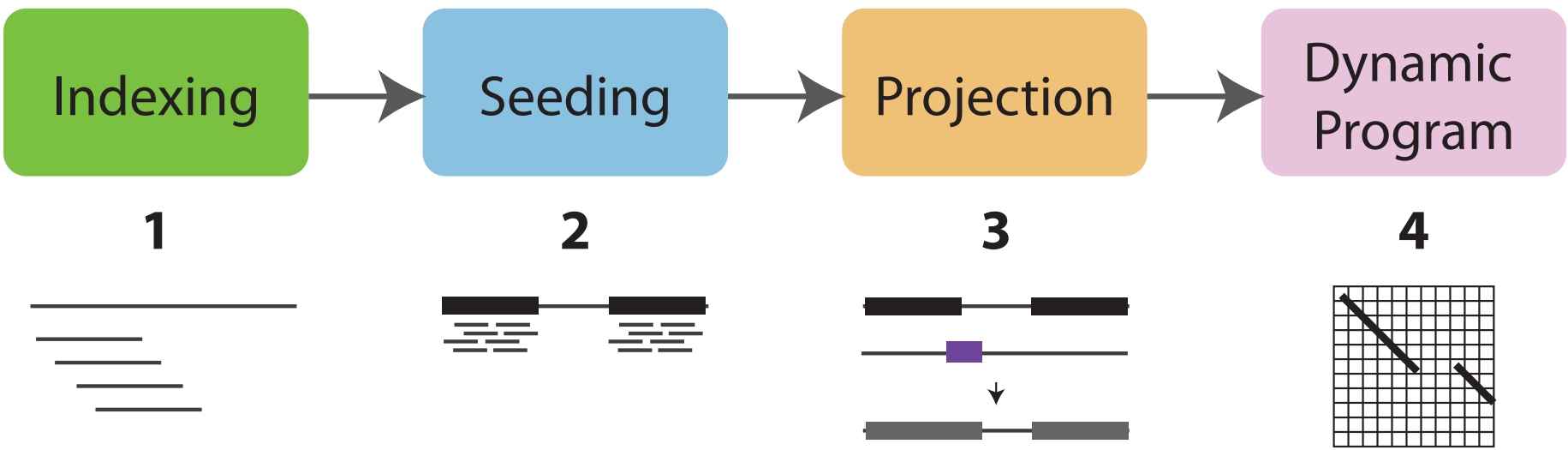
## General Workflow

- general seed-and-extend strategy
- following four basic steps

# RNA-seq Read Alignments with PALMapper

## General Workflow

- general seed-and-extend strategy
- following four basic steps

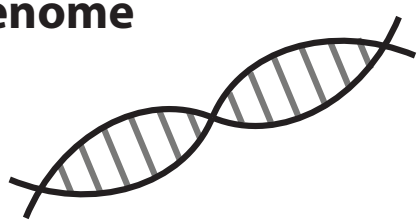


# RNA-seq Read Alignments with PALMapper

## Index and Seeding (normal case)

- Indexing genome with $k$-mer based index or bwt-based index
- Lookup read $k$-mers in the genome index
- Generate long and short seed regions for further alignment

**Genome**



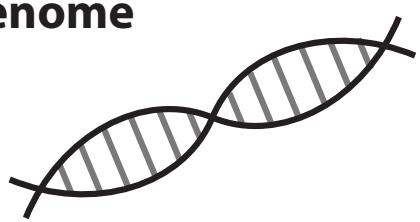
**HTS Reads**

# RNA-seq Read Alignments with PALMapper

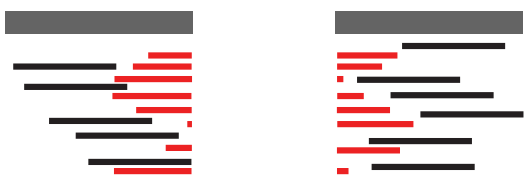
## Index and Seeding (normal case)

- Indexing genome with $k$-mer based index or bwt-based index
- Lookup read $k$-mers in the genome index
- Generate long and short seed regions for further alignment

**Genome**



**HTS Reads**



# RNA-seq Read Alignments with PALMapper

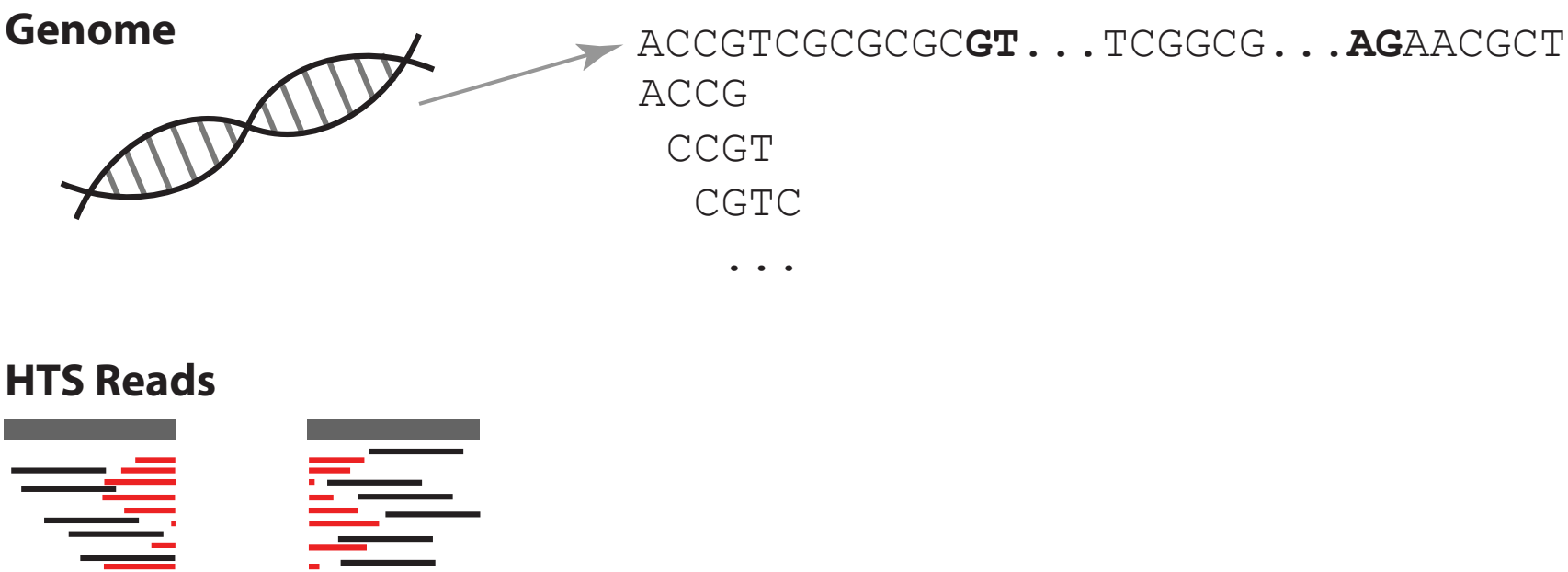
## Index and Seeding (normal case)

- Indexing genome with *k*-mer based index or bwt-based index
- Lookup read *k*-mers in the genome index
- Generate long and short seed regions for further alignment

# RNA-seq Read Alignments with PALMapper

## Index and Seeding (normal case)

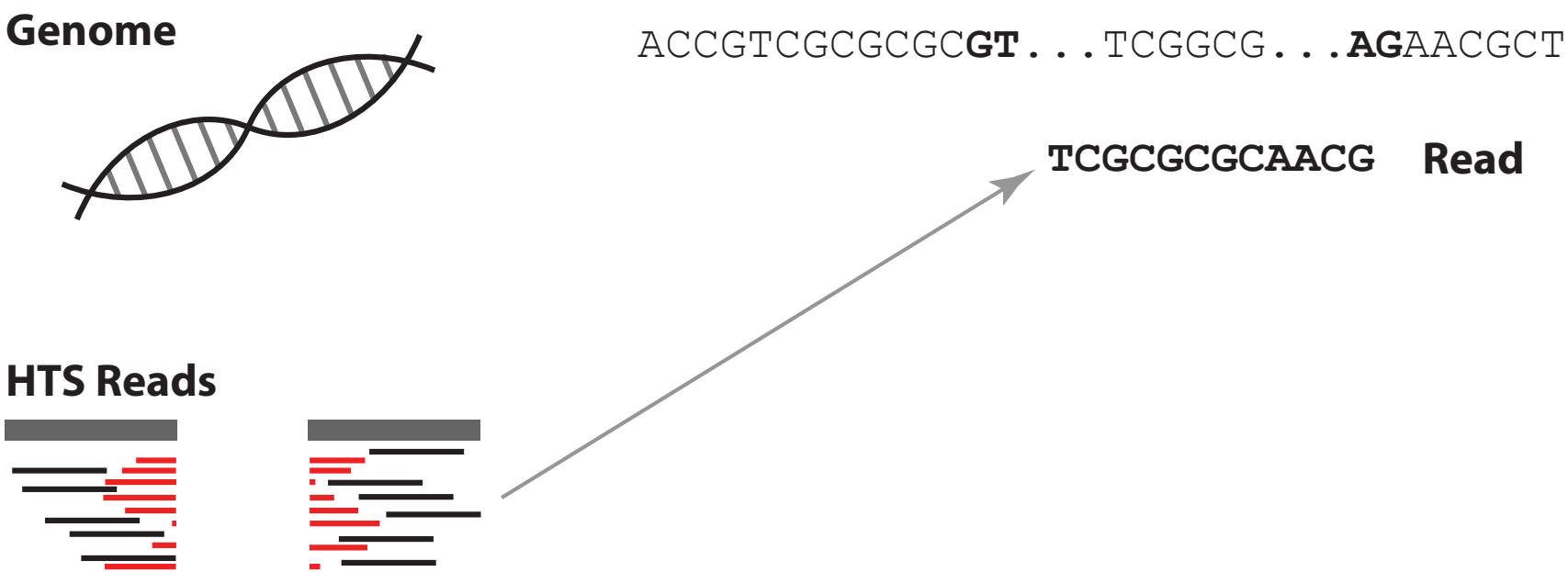
- Indexing genome with *k*-mer based index or bwt-based index
- Lookup read *k*-mers in the genome index
- Generate long and short seed regions for further alignment



# RNA-seq Read Alignments with PALMapper

## Index and Seeding (normal case)

- Indexing genome with *k*-mer based index or bwt-based index
- Lookup read *k*-mers in the genome index
- Generate long and short seed regions for further alignment

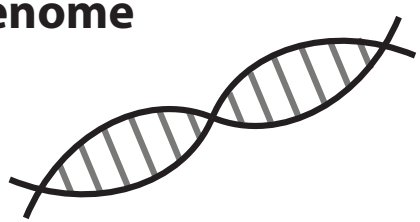


# RNA-seq Read Alignments with PALMapper

## Index and Seeding (normal case)

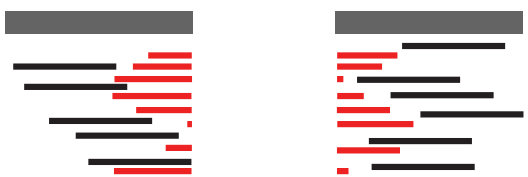
- Indexing genome with *k*-mer based index or bwt-based index
- Lookup read *k*-mers in the genome index
- Generate long and short seed regions for further alignment

**Genome**



ACCGTCGCGCGC**GT**...TCGGCG...**AG**AACGCT

**HTS Reads**

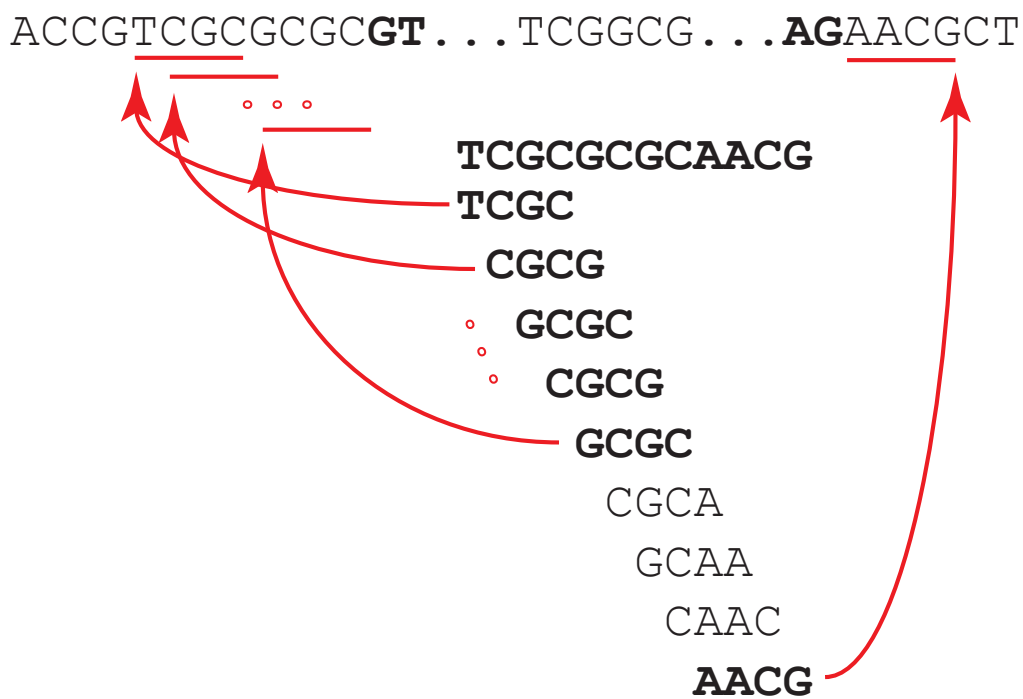


**TCGCGCGCAACG** **Read**
**TCGC**
**CGCG**
**GCGC**
**CGCG**
**GCGC**
CGCA
GCAA
CAAC
**AACG**

# RNA-seq Read Alignments with PALMapper

## Index and Seeding (normal case)

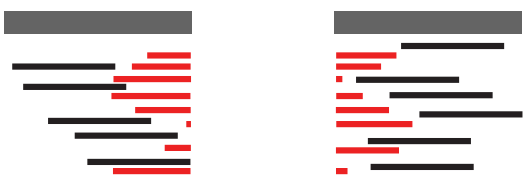
- Indexing genome with *k*-mer based index or bwt-based index
- Lookup read *k*-mers in the genome index
- Generate long and short seed regions for further alignment

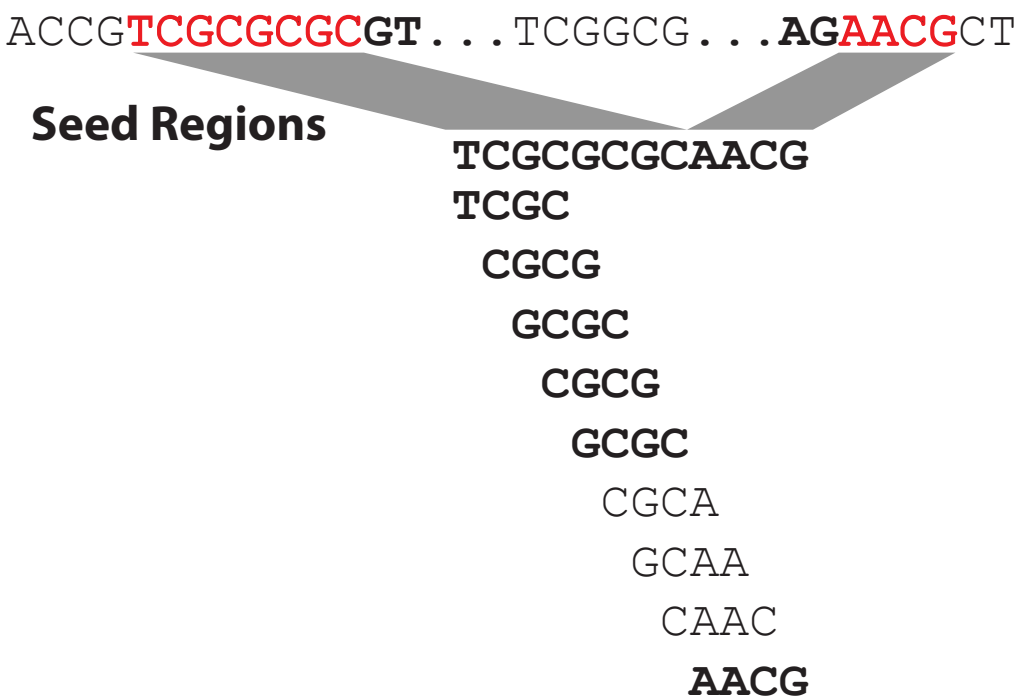



# RNA-seq Read Alignments with PALMapper

## Index and Seeding (normal case)

- Indexing genome with *k*-mer based index or bwt-based index
- Lookup read *k*-mers in the genome index
- Generate long and short seed regions for further alignment

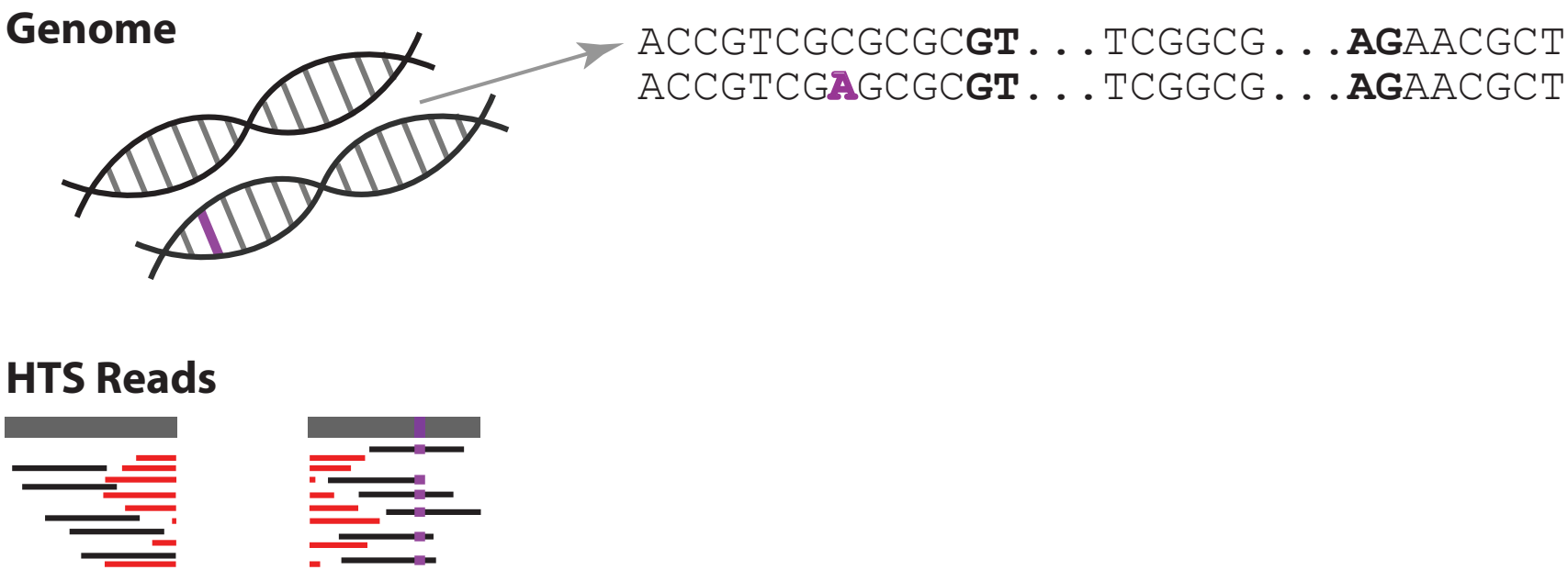



# Variation-Aware Indexing

## Indexing (variant case)

- Include available variation into index (dbSNP, COSMIC, ENCODE, 1000 Genomes . . . )
- Principle for querying the index stays the same



# Variation-Aware Indexing

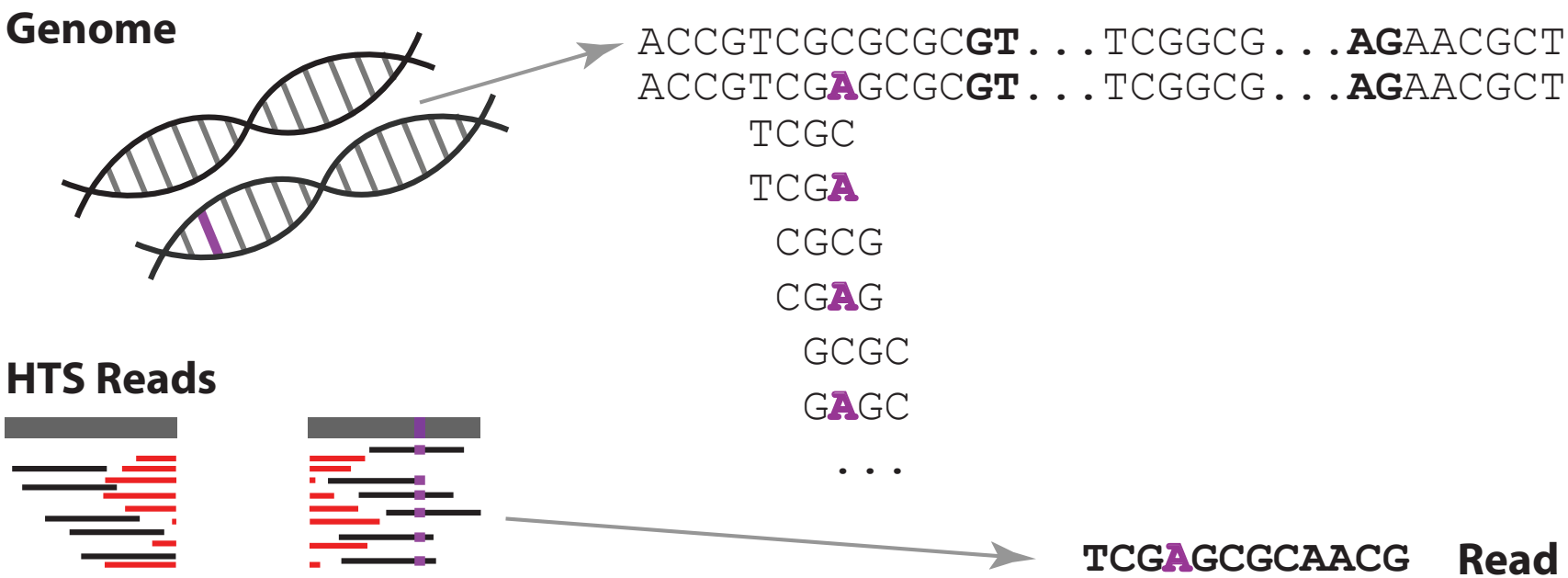
## Indexing (variant case)

- Include available variation into index (dbSNP, COSMIC, ENCODE, 1000 Genomes . . . )
- Principle for querying the index stays the same

# Variation-Aware Indexing

## Indexing (variant case)

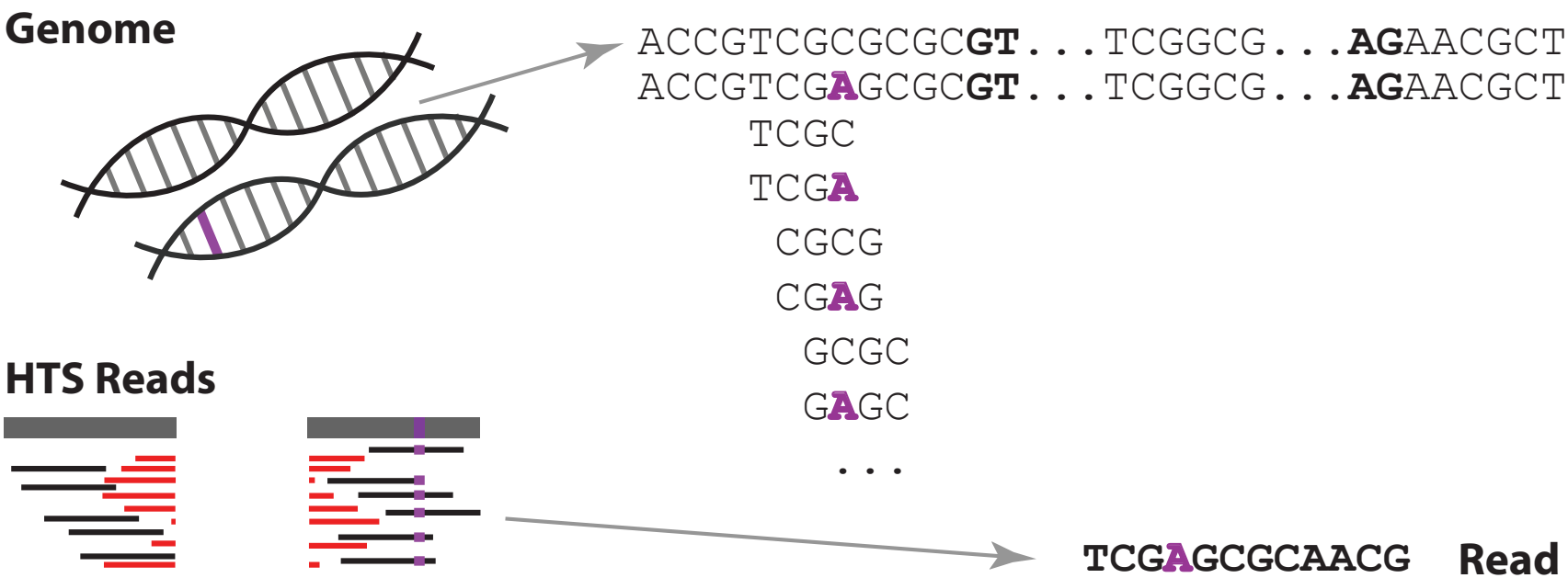
- Include available variation into index (dbSNP, COSMIC, ENCODE, 1000 Genomes . . . )
- Principle for querying the index stays the same



# Variation-Aware Indexing

## Indexing (variant case)

- Include available variation into index (dbSNP, COSMIC, ENCODE, 1000 Genomes . . . )
- **Principle for querying the index stays the same**



# Variation-Aware Indexing

## Additional to SNPs PALMapper can account for

- Insertions
- Deletions
- Combinations of variants (incl. unbalanced substitutions)

**Insertions**

```
ACCG---TCGCGAGCGT...TCGGCG...AGAACGCT
ACCGCGTTCGCGAGCGT...TCGGCG...AGAACGCT
   ...
    G---TCG
    GCGT
     CGTT
      GTTC
       TTCG
         ...
```

# Variation-Aware Indexing

## Additional to SNPs PALMapper can account for

- Insertions
- Deletions
- Combinations of variants (incl. unbalanced substitutions)

**Deletions**

```
ACCGTCGCGAGCGT...TCGGCG...AGAACGCT
ACCG--GCGAGCGT...TCGGCG...AGAACGCT
 CCG--G
 CCGT
  CG--GC
  CGTC
   G--GCG
   GTCG
     ...
```

# Variation-Aware Indexing

## Additional to SNPs PALMapper can account for

- Insertions
- Deletions
- Combinations of variants (incl. unbalanced substitutions)

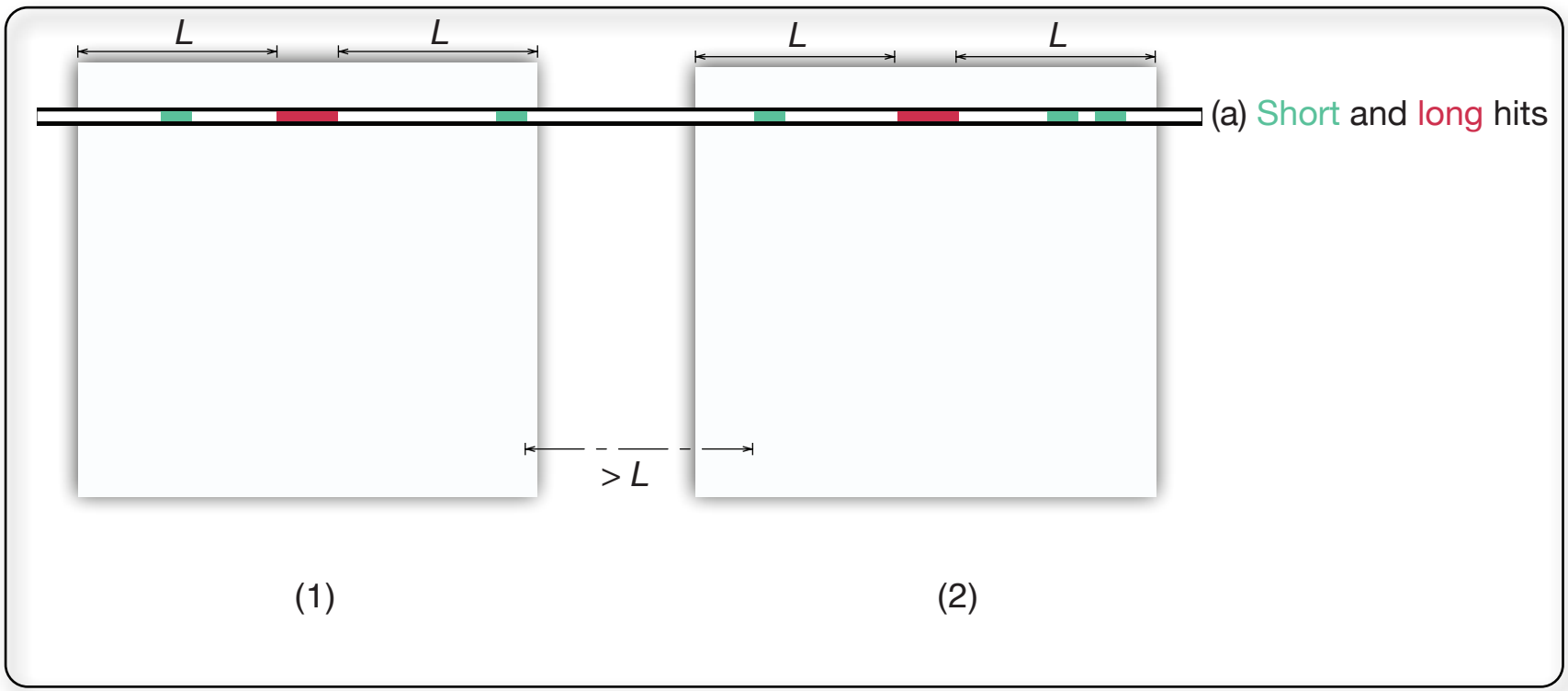
**Combinations**

```
ACCGTCGCGAGCGT...TCGGCG...AGAACGCT
ACCGTTGAGAGCGT...TCGGCG...AGAACGCT
    TCGC
    TTGC
    TCGA
    TTGA
     ...
```

# Genomic Sequence Reduction

## Projection to Pseudo Chromosomes

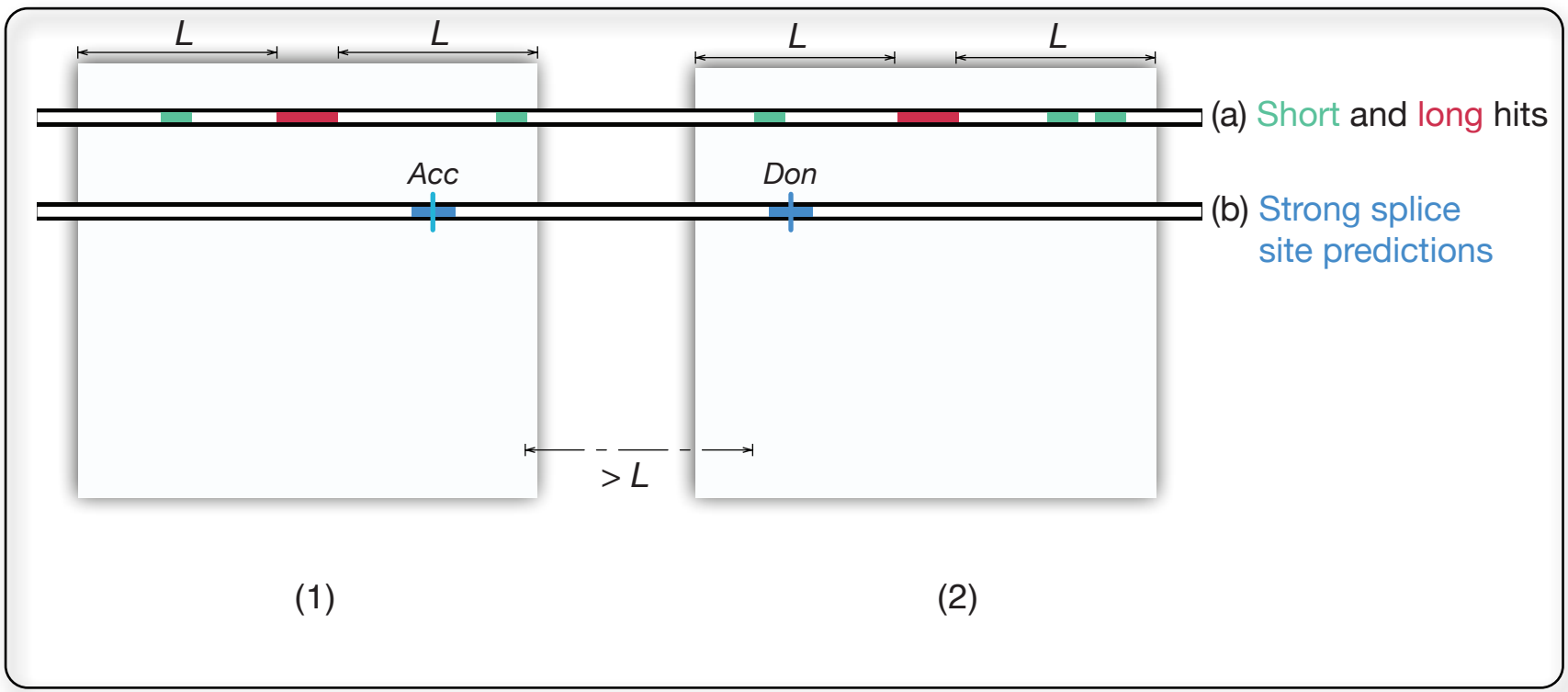
- Rank and combine seed regions from previous step
- Integrate with splice site prediction and coverage information
- Project to genome and form *Pseudo Chromosome Sequences*



# Genomic Sequence Reduction

## Projection to Pseudo Chromosomes

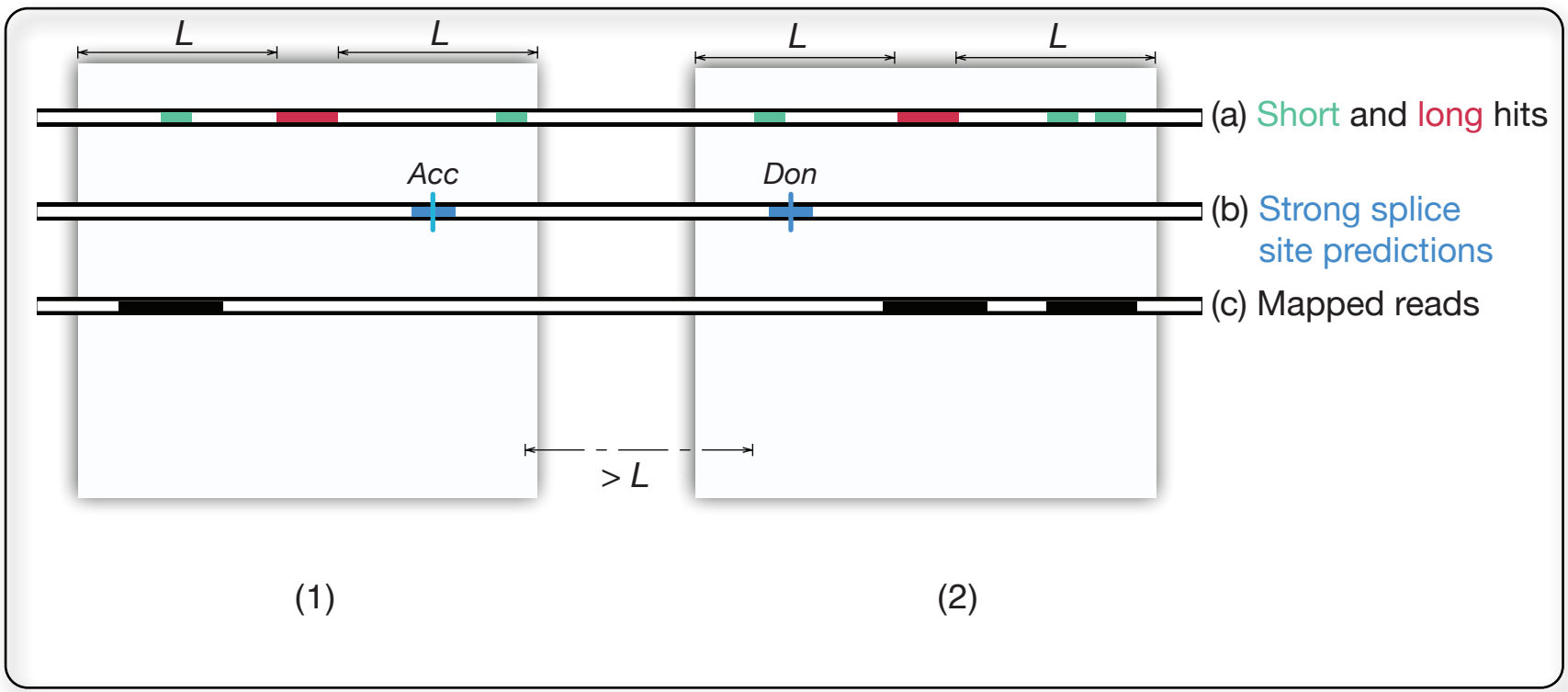
- Rank and combine seed regions from previous step
- Integrate with splice site prediction and coverage information
- Project to genome and form *Pseudo Chromosome Sequences*



# Genomic Sequence Reduction

## Projection to Pseudo Chromosomes

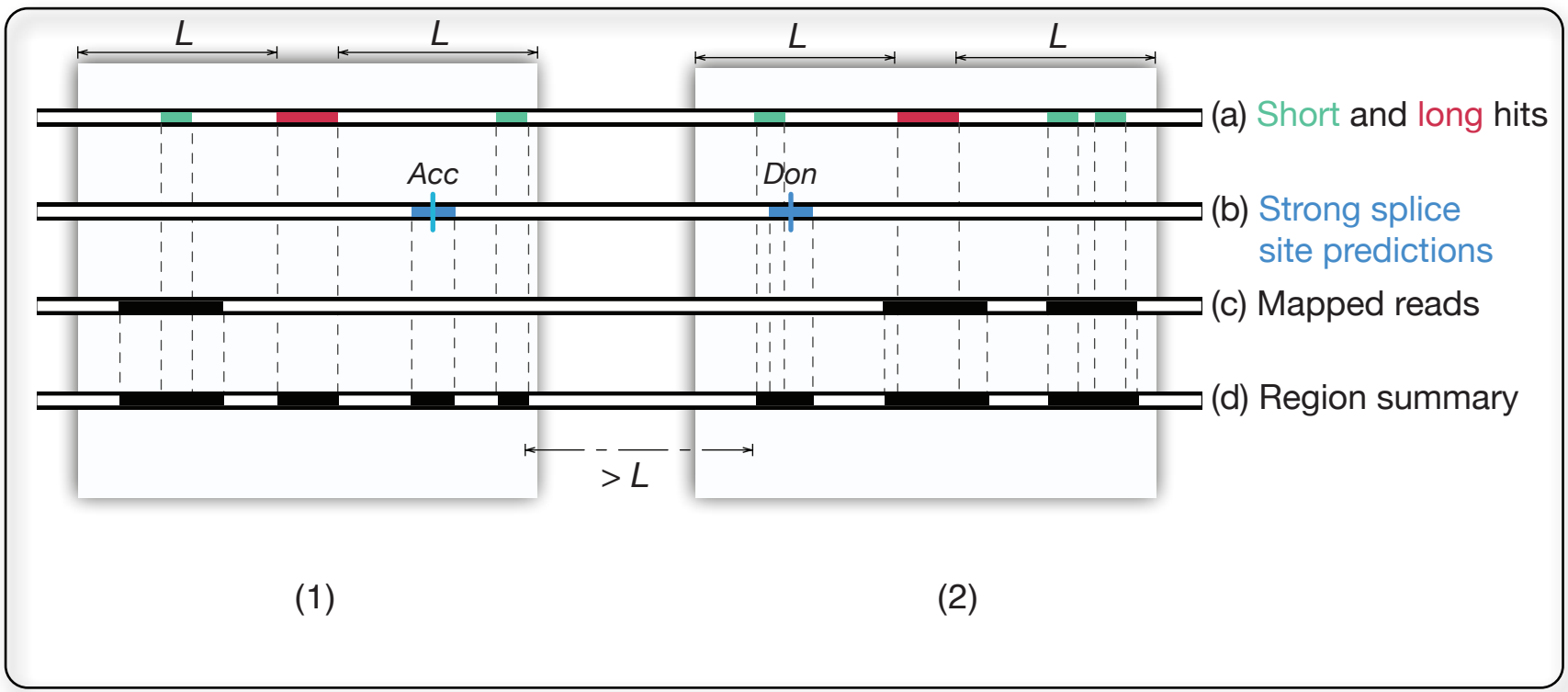
- Rank and combine seed regions from previous step
- Integrate with splice site prediction and coverage information
- Project to genome and form *Pseudo Chromosome Sequences*



# Genomic Sequence Reduction

## Projection to Pseudo Chromosomes

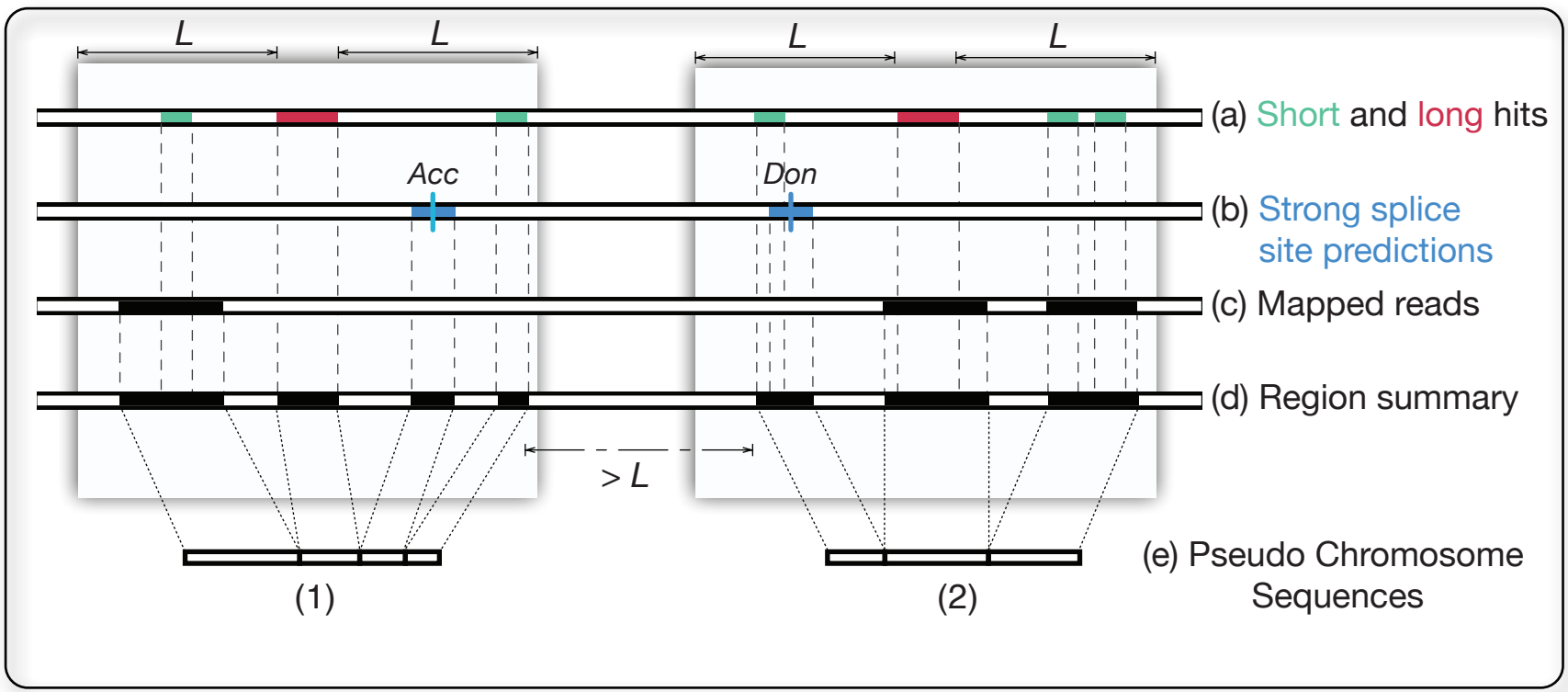
- Rank and combine seed regions from previous step
- Integrate with splice site prediction and coverage information
- Project to genome and form *Pseudo Chromosome Sequences*



# Genomic Sequence Reduction

## Projection to Pseudo Chromosomes

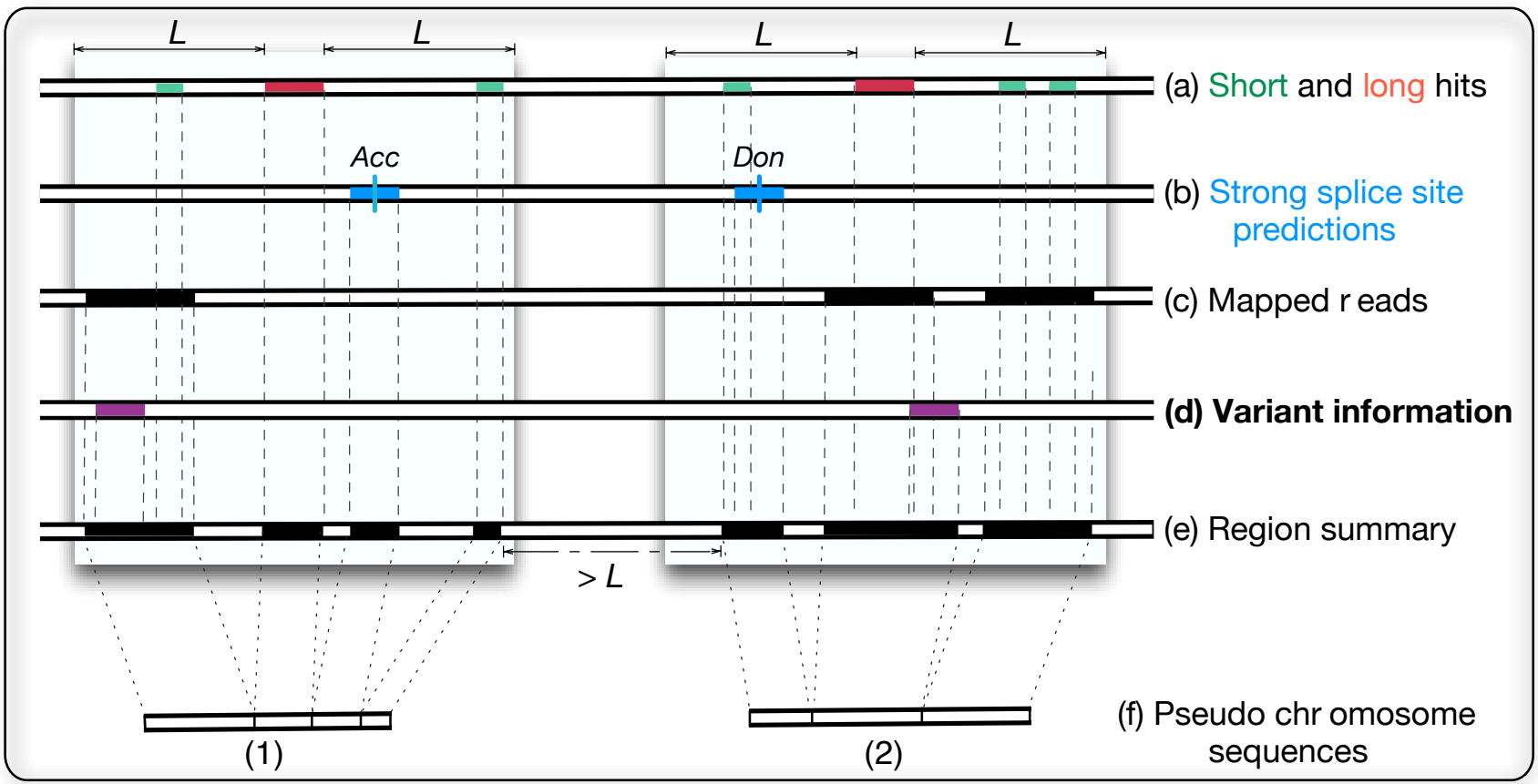
- Rank and combine seed regions from previous step
- Integrate with splice site prediction and coverage information
- Project to genome and form *Pseudo Chromosome Sequences*



# Genomic Sequence Reduction with Variation

## Projection to Pseudo Chromosomes (variant case)

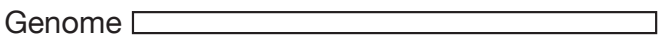
- Include known sequence variants to form combination of pseudo chromosomes



# Genomic Sequence Reduction with Variation

## Dynamic Program Alignment Against Graph

- Integrate substitutions as IUPAC codes for ambiguous bases
- Integrate deletions as jumps
- Integrate insertions as jumps
- Collapse all variations into graph-like structure
- Align against variation graph

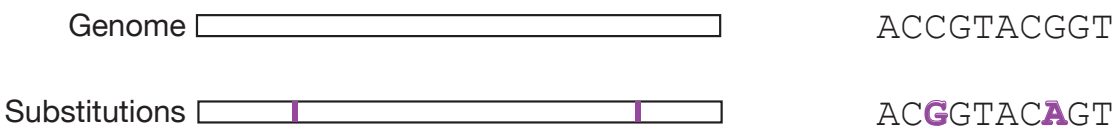



ACCGTACGGT

# Genomic Sequence Reduction with Variation

## Dynamic Program Alignment Against Graph

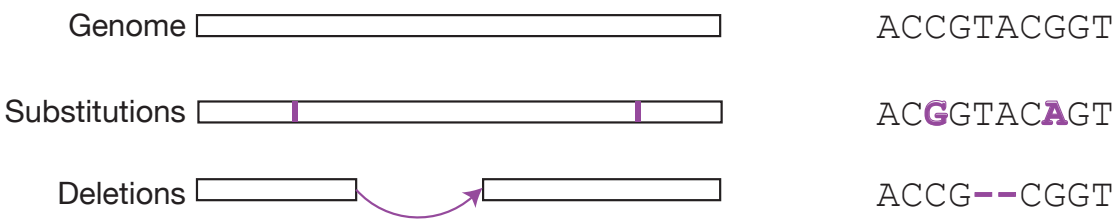
- Integrate substitutions as IUPAC codes for ambiguous bases
- Integrate deletions as jumps
- Integrate insertions as jumps
- Collapse all variations into graph-like structure
- Align against variation graph



# Genomic Sequence Reduction with Variation

## Dynamic Program Alignment Against Graph

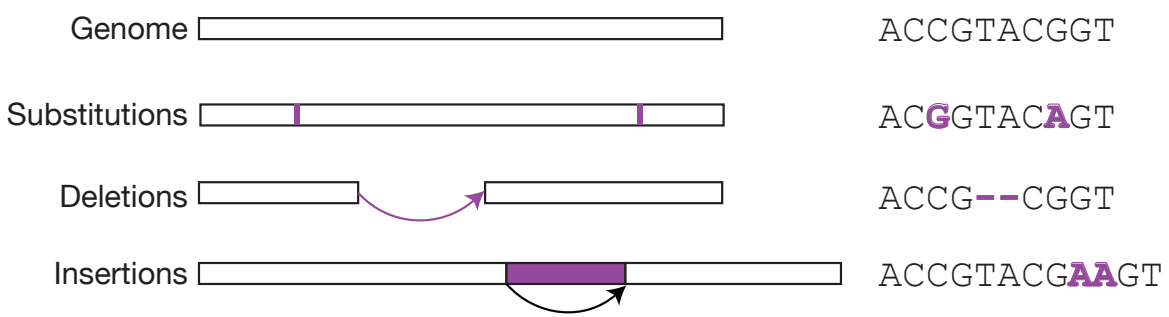
- Integrate substitutions as IUPAC codes for ambiguous bases
- Integrate deletions as jumps
- Integrate insertions as jumps
- Collapse all variations into graph-like structure
- Align against variation graph



# Genomic Sequence Reduction with Variation

## Dynamic Program Alignment Against Graph

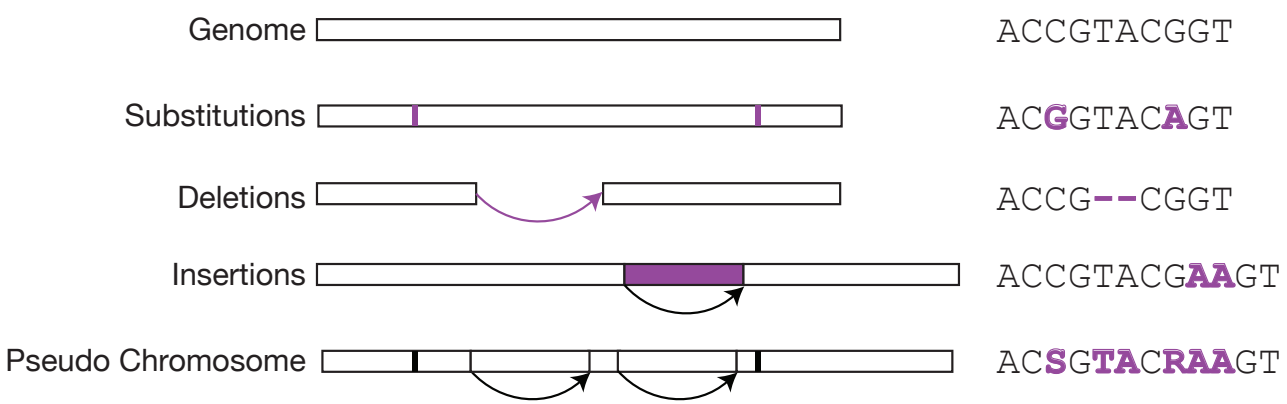
- Integrate substitutions as IUPAC codes for ambiguous bases
- Integrate deletions as jumps
- Integrate insertions as jumps
- Collapse all variations into graph-like structure
- Align against variation graph



# Genomic Sequence Reduction with Variation

## Dynamic Program Alignment Against Graph

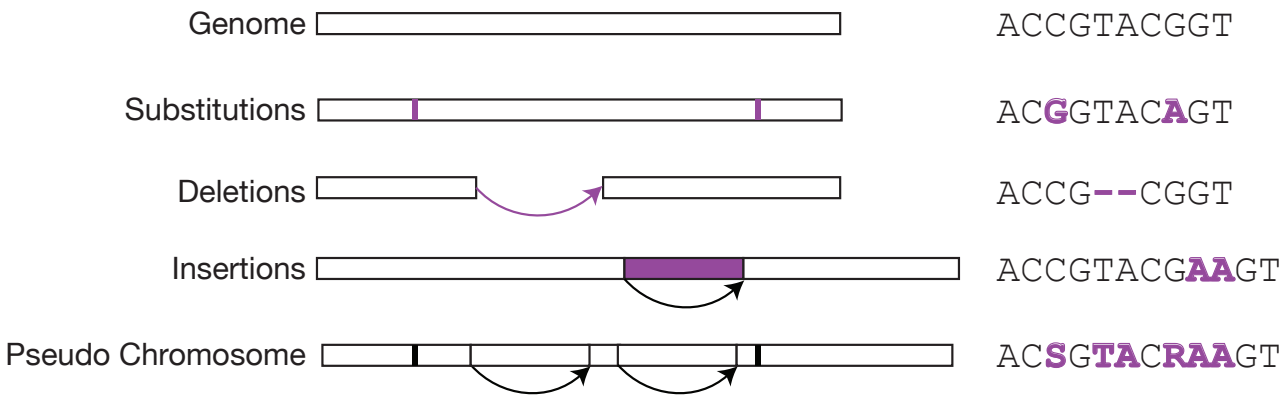
- Integrate substitutions as IUPAC codes for ambiguous bases
- Integrate deletions as jumps
- Integrate insertions as jumps
- Collapse all variations into graph-like structure
- Align against variation graph

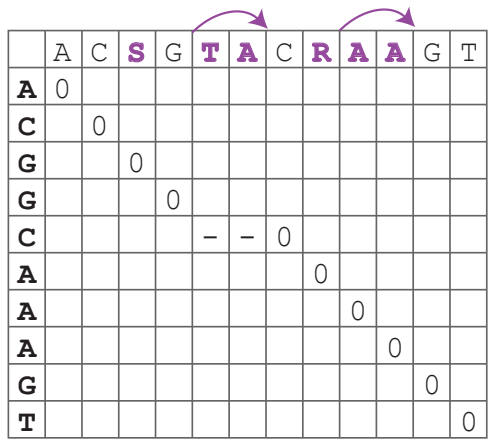


# Genomic Sequence Reduction with Variation

## Dynamic Program Alignment Against Graph

- Integrate substitutions as IUPAC codes for ambiguous bases
- Integrate deletions as jumps
- Integrate insertions as jumps
- Collapse all variations into graph-like structure
- Align against variation graph



# Dynamic Program (DP)

The inferred *seed region* guides a DP-based alignment algorithm:


| | gap | A | C | G | T | N |
|---|---|---|---|---|---|---|
| gap | 0.33 | $f_{-A}(\cdot)$ | $f_{-C}(\cdot)$ | $f_{-G}(\cdot)$ | $f_{-T}(\cdot)$ | $f_{-N}(\cdot)$ |
| A | 0.31 | $f_{AA}(\cdot)$ | $f_{AC}(\cdot)$ | $f_{AG}(\cdot)$ | $f_{AT}(\cdot)$ | $f_{AN}(\cdot)$ |
| C | 0.44 | $f_{CA}(\cdot)$ | $f_{CC}(\cdot)$ | $f_{CG}(\cdot)$ | $f_{CT}(\cdot)$ | $f_{CN}(\cdot)$ |
| G | 0.13 | $f_{GA}(\cdot)$ | $f_{GC}(\cdot)$ | $f_{GG}(\cdot)$ | $f_{GT}(\cdot)$ | $f_{GN}(\cdot)$ |
| T | 0.55 | $f_{TA}(\cdot)$ | $f_{TC}(\cdot)$ | $f_{TG}(\cdot)$ | $f_{TT}(\cdot)$ | $f_{TN}(\cdot)$ |
| N | 0.12 | $f_{NA}(\cdot)$ | $f_{NC}(\cdot)$ | $f_{NG}(\cdot)$ | $f_{NT}(\cdot)$ | $f_{NN}(\cdot)$ |

QPALMA Scores are functions of:

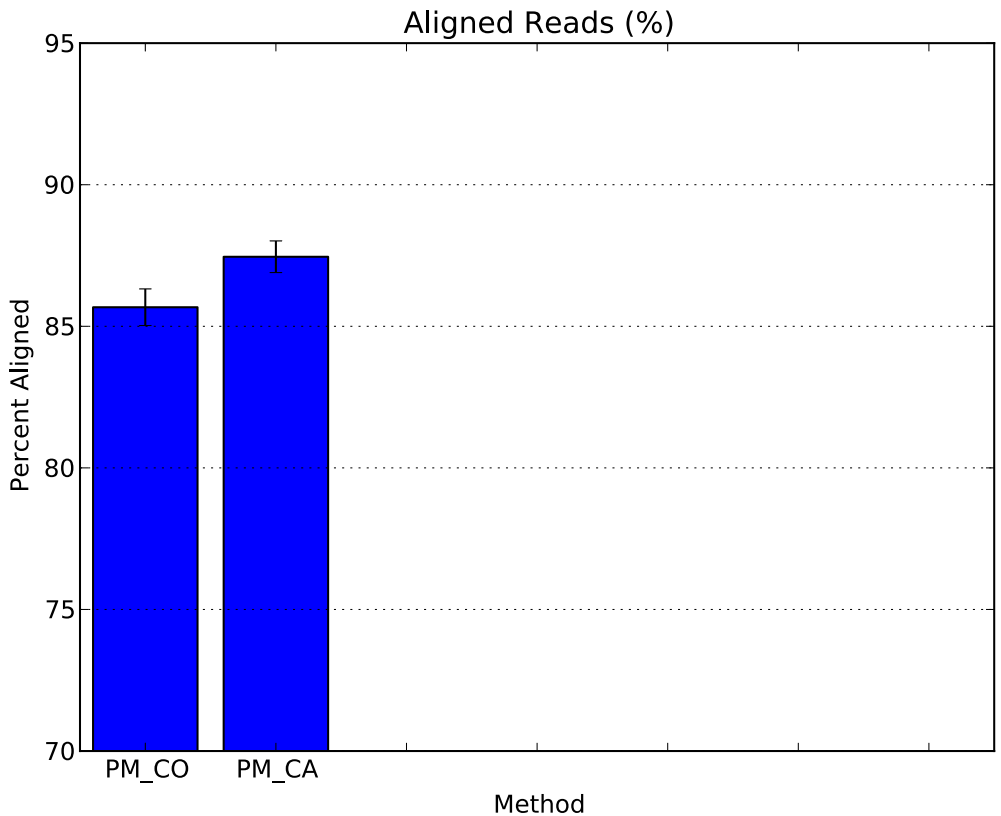
- splice site strengths
- intron length
- quality information
- error profile

| Semi-Global | Banded | Spliced | Variation-aware |
|---|---|---|---|
| Align whole read to part of pseudo chromosome | Limit # of gaps from perfect alignment | Allow long gaps corresponding to introns | Account for genomic variations |

# Alignment Sensitivity

## Variant alignments cause gain in sensitivity

- RNA-Seq of *A. thaliana* strain Can-0 aligned to reference Col-0
- use variation during seeding and extension improves sensitivity



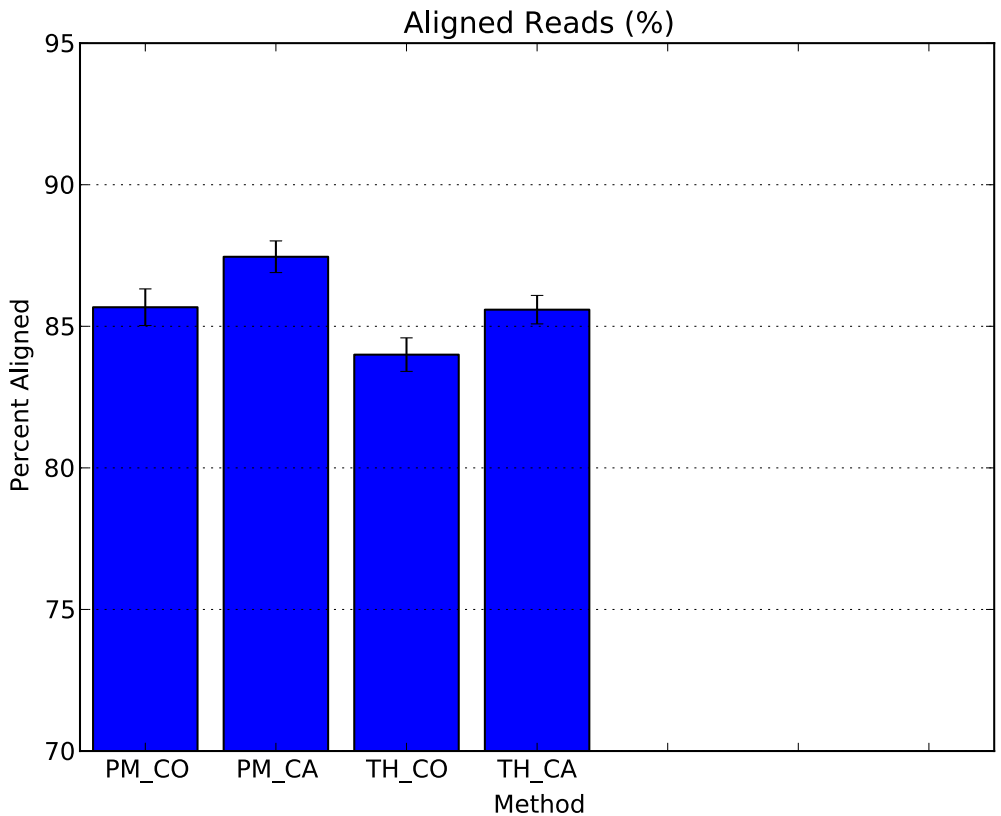

Legend:

| | |
|---|---|
| PM | PALMapper |
| TH | TopHat |
| CO | aligned to Col-0 |
| CA | aligned to Can-0 |
| vi | variant indexing |
| v | variant alignment |
| vvi | v and vi together |

# Alignment Sensitivity

## Variant alignments cause gain in sensitivity

- RNA-Seq of *A. thaliana* strain Can-0 aligned to reference Col-0
- use variation during seeding and extension improves sensitivity



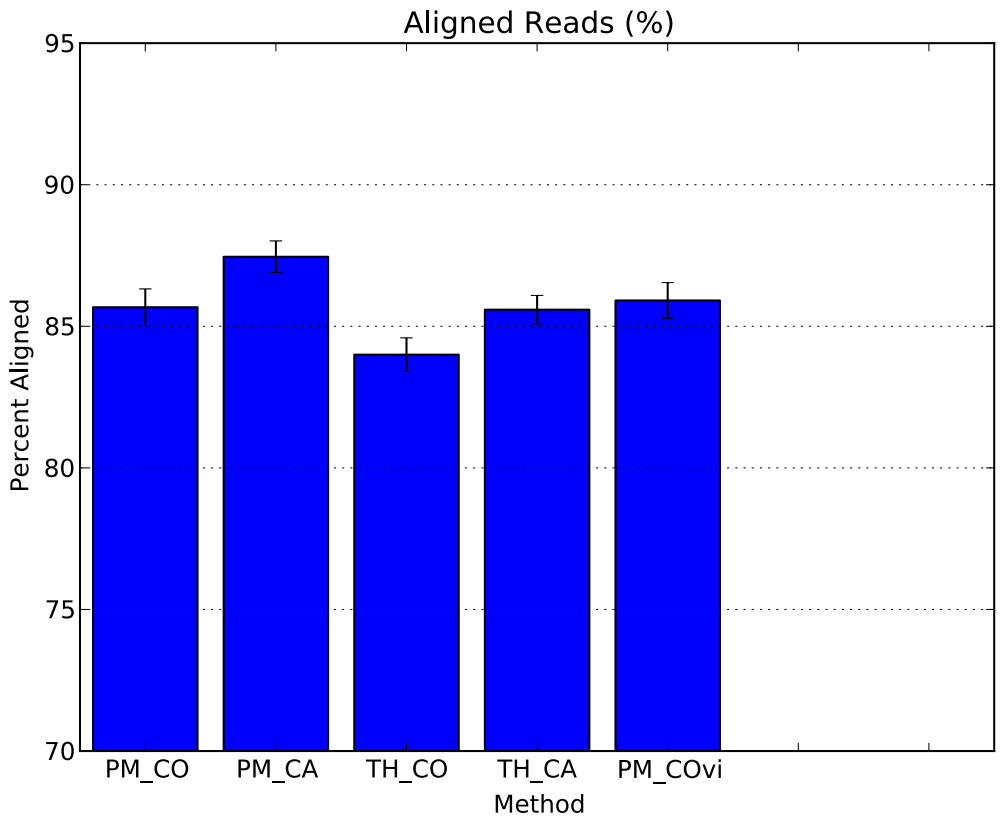

Legend:

| | |
|---|---|
| PM | PALMapper |
| TH | TopHat |
| CO | aligned to Col-0 |
| CA | aligned to Can-0 |
| vi | variant indexing |
| v | variant alignment |
| vvi | v and vi together |

# Alignment Sensitivity

## Variant alignments cause gain in sensitivity

- RNA-Seq of *A. thaliana* strain Can-0 aligned to reference Col-0
- use variation during seeding and extension improves sensitivity



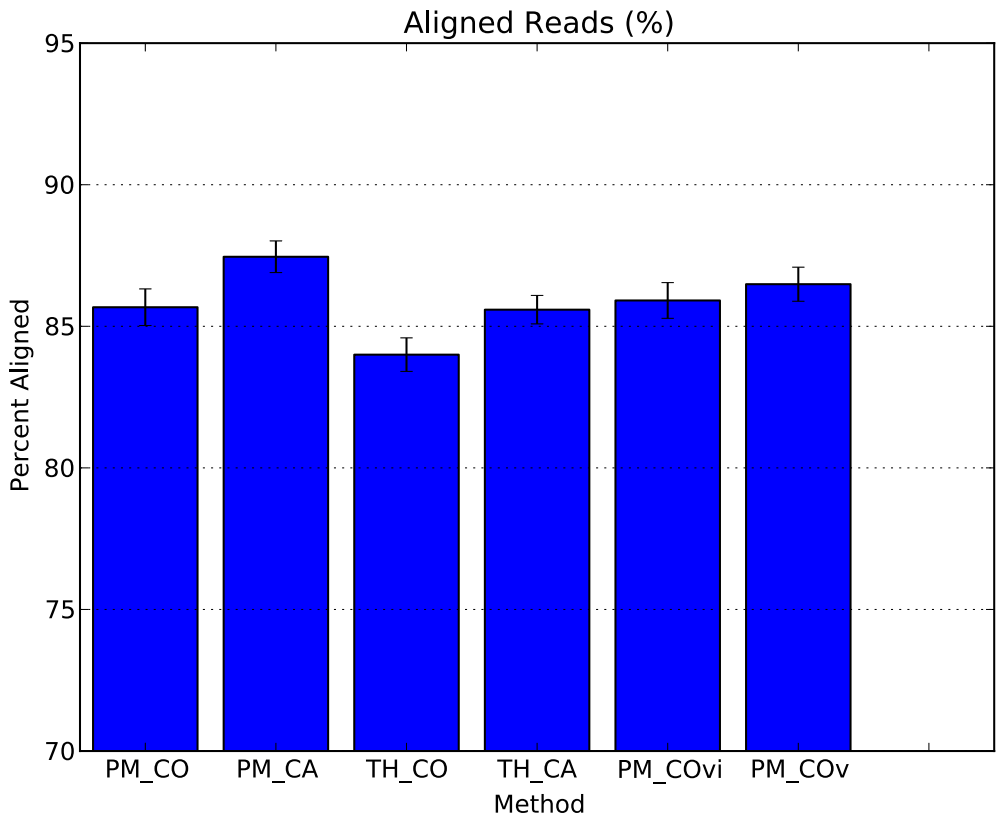

Legend:

| | |
|---|---|
| PM | PALMapper |
| TH | TopHat |
| CO | aligned to Col-0 |
| CA | aligned to Can-0 |
| vi | variant indexing |
| v | variant alignment |
| vvi | v and vi together |

# Alignment Sensitivity

## Variant alignments cause gain in sensitivity

- RNA-Seq of *A. thaliana* strain Can-0 aligned to reference Col-0
- use variation during seeding and extension improves sensitivity



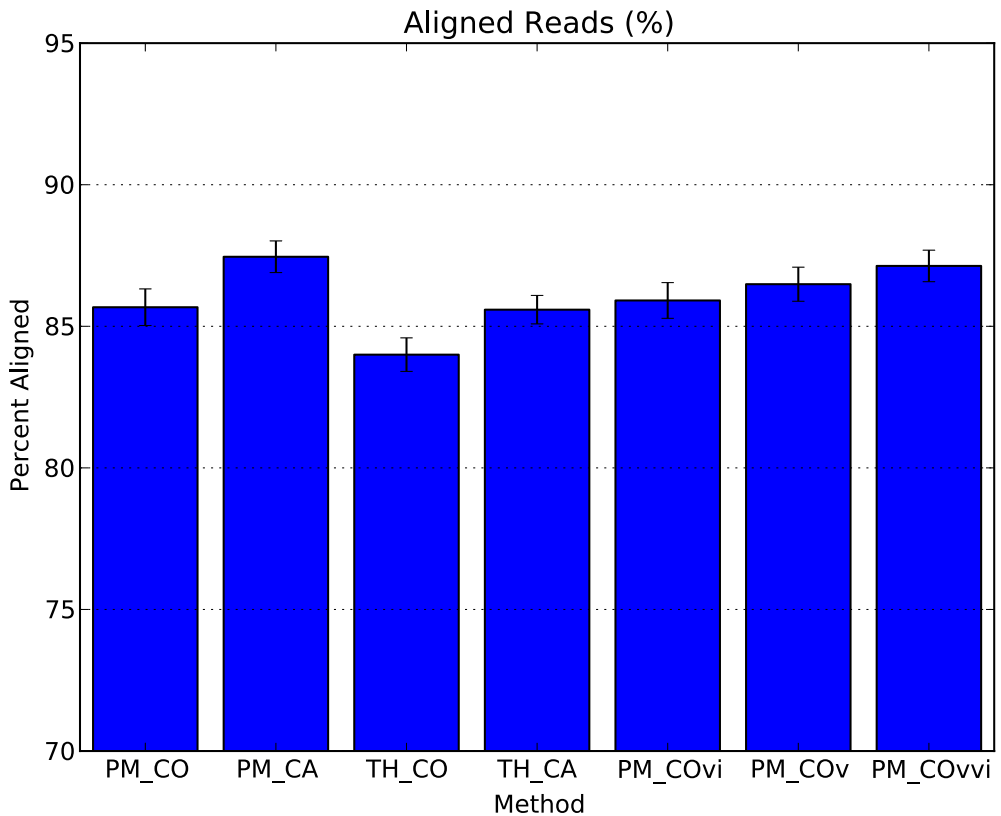

Legend:

| | |
|---|---|
| PM | PALMapper |
| TH | TopHat |
| CO | aligned to Col-0 |
| CA | aligned to Can-0 |
| vi | variant indexing |
| v | variant alignment |
| vvi | v and vi together |

# Alignment Sensitivity

## Variant alignments cause gain in sensitivity

- RNA-Seq of *A. thaliana* strain Can-0 aligned to reference Col-0
- use variation during seeding and extension improves sensitivity



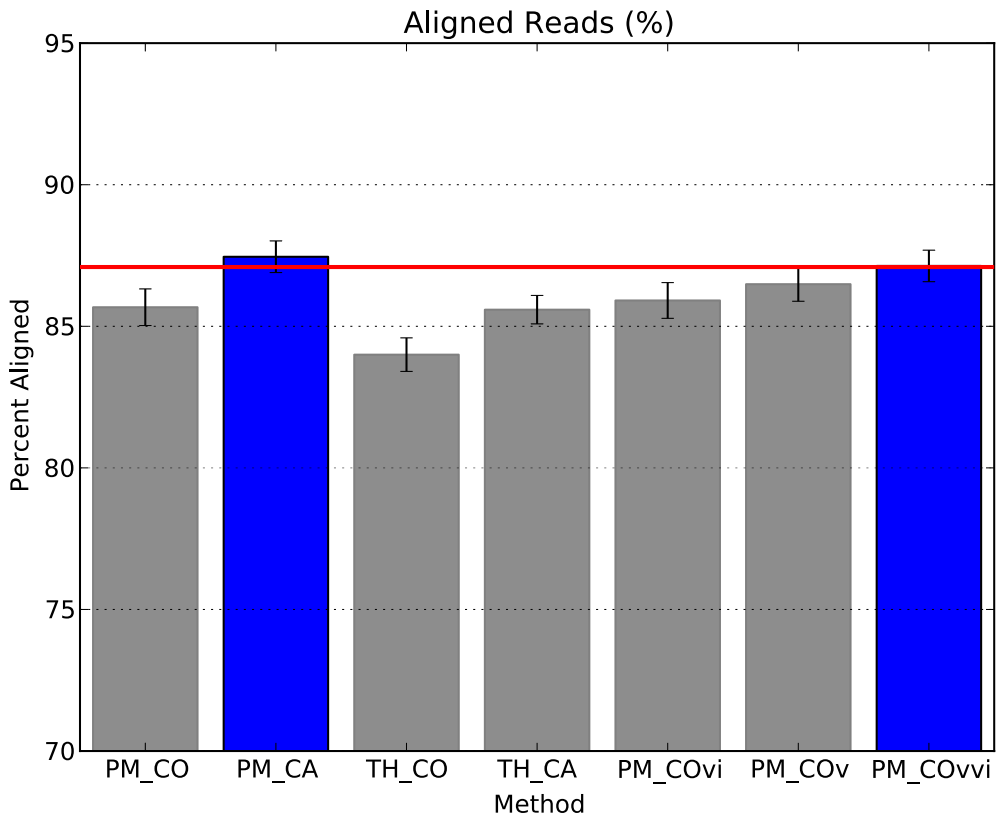

Legend:

| | |
|---|---|
| PM | PALMapper |
| TH | TopHat |
| CO | aligned to Col-0 |
| CA | aligned to Can-0 |
| vi | variant indexing |
| v | variant alignment |
| vvi | v and vi together |

# Alignment Sensitivity

## Variant alignments cause gain in sensitivity

- RNA-Seq of *A. thaliana* strain Can-0 aligned to reference Col-0
- use variation during seeding and extension improves sensitivity




Legend:

| | |
|---|---|
| PM | PALMapper |
| TH | TopHat |
| CO | aligned to Col-0 |
| CA | aligned to Can-0 |
| vi | variant indexing |
| v | variant alignment |
| vvi | v and vi together |

# Running Time

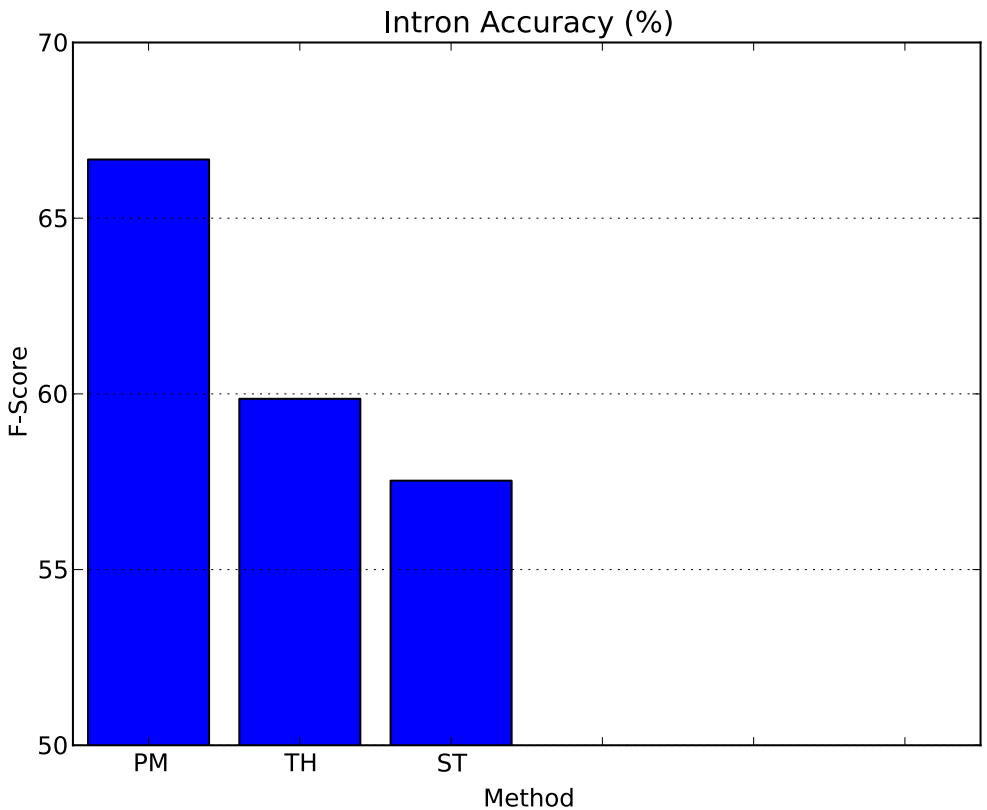
## Running time and Scaling

- 2M reads / h / core on human genome
- ca 1M reads / h /core with variants (complexity dependent)
- almost linear scaling with number of cores

# Accuracy of Intron Prediction

## Comparison of predicted introns to annotation

- PALMapper outperforms TopHat and STAR in terms of accuracy
- Variation-aware alignment increases performance



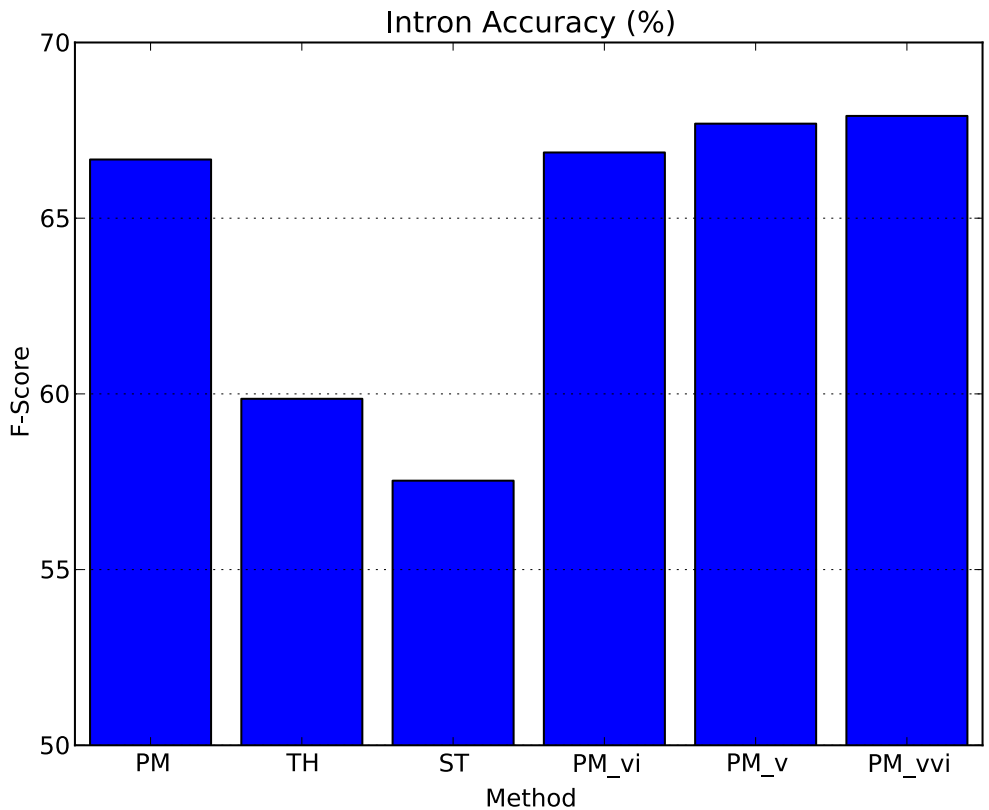

Legend:

| | |
|---|---|
| PM | PALMapper |
| TH | TopHat |
| ST | STAR |
| vi | variant indexing |
| v | variant alignment |
| vvi | v and vi together |

# Accuracy of Intron Prediction

## Comparison of predicted introns to annotation

- PALMapper outperforms TopHat and STAR in terms of accuracy
- Variation-aware alignment increases performance



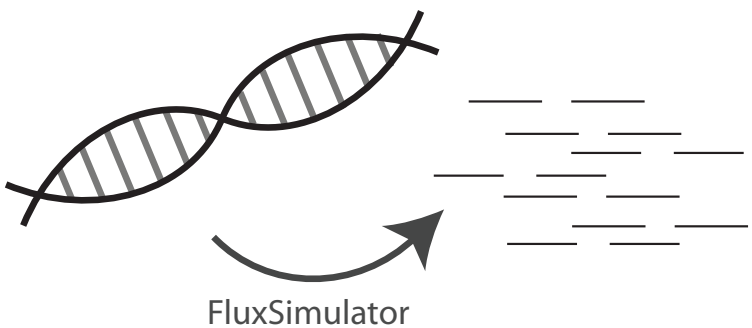

Legend:

| | |
|---|---|
| PM | PALMapper |
| TH | TopHat |
| ST | STAR |
| vi | variant indexing |
| v | variant alignment |
| vvi | v and vi together |

# Artificial Read Set

## Construct read set to measure allelic imbalance

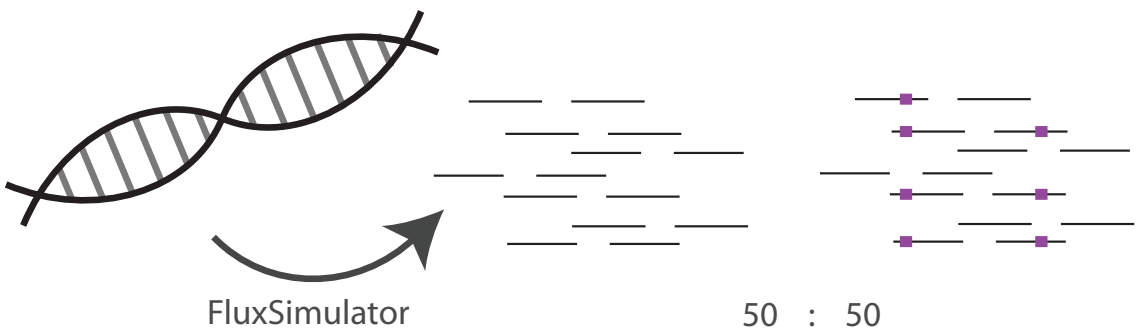
- Use FluxSimulator to generate artificial reads (10M from 5K genes)
- Duplicate and mutate read set
- Re-align to original genome
- Expectation: perfect allelic balance!



# Artificial Read Set

## Construct read set to measure allelic imbalance

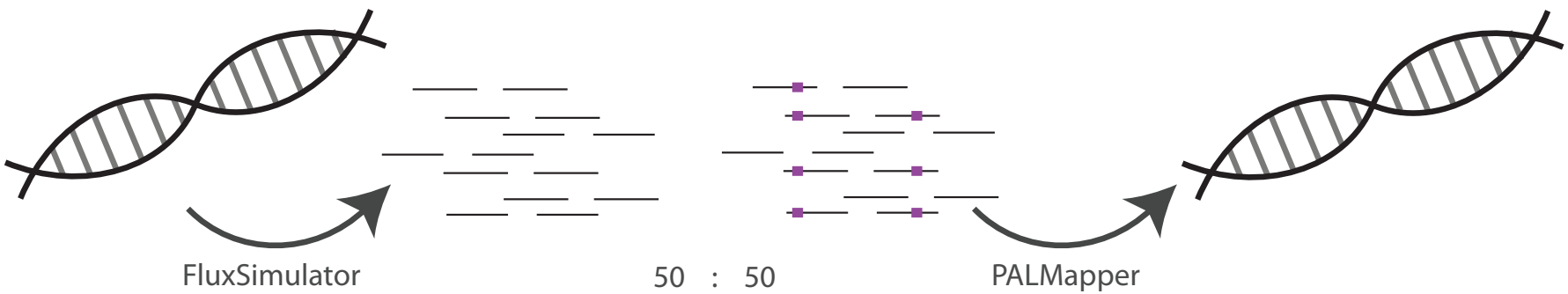
- Use FluxSimulator to generate artificial reads (10M from 5K genes)
- Duplicate and mutate read set
- Re-align to original genome
- Expectation: perfect allelic balance!



# Artificial Read Set

## Construct read set to measure allelic imbalance

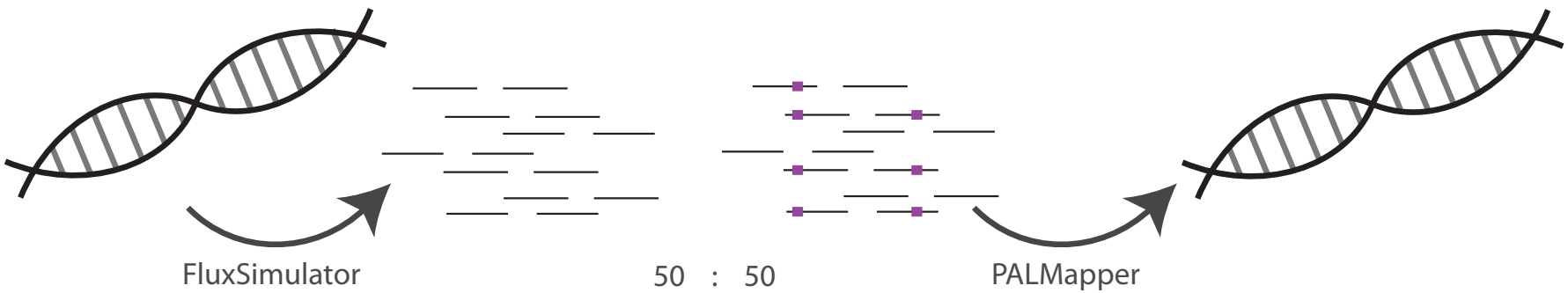
- Use FluxSimulator to generate artificial reads (10M from 5K genes)
- Duplicate and mutate read set
- Re-align to original genome
- Expectation: perfect allelic balance!



# Artificial Read Set

## Construct read set to measure allelic imbalance

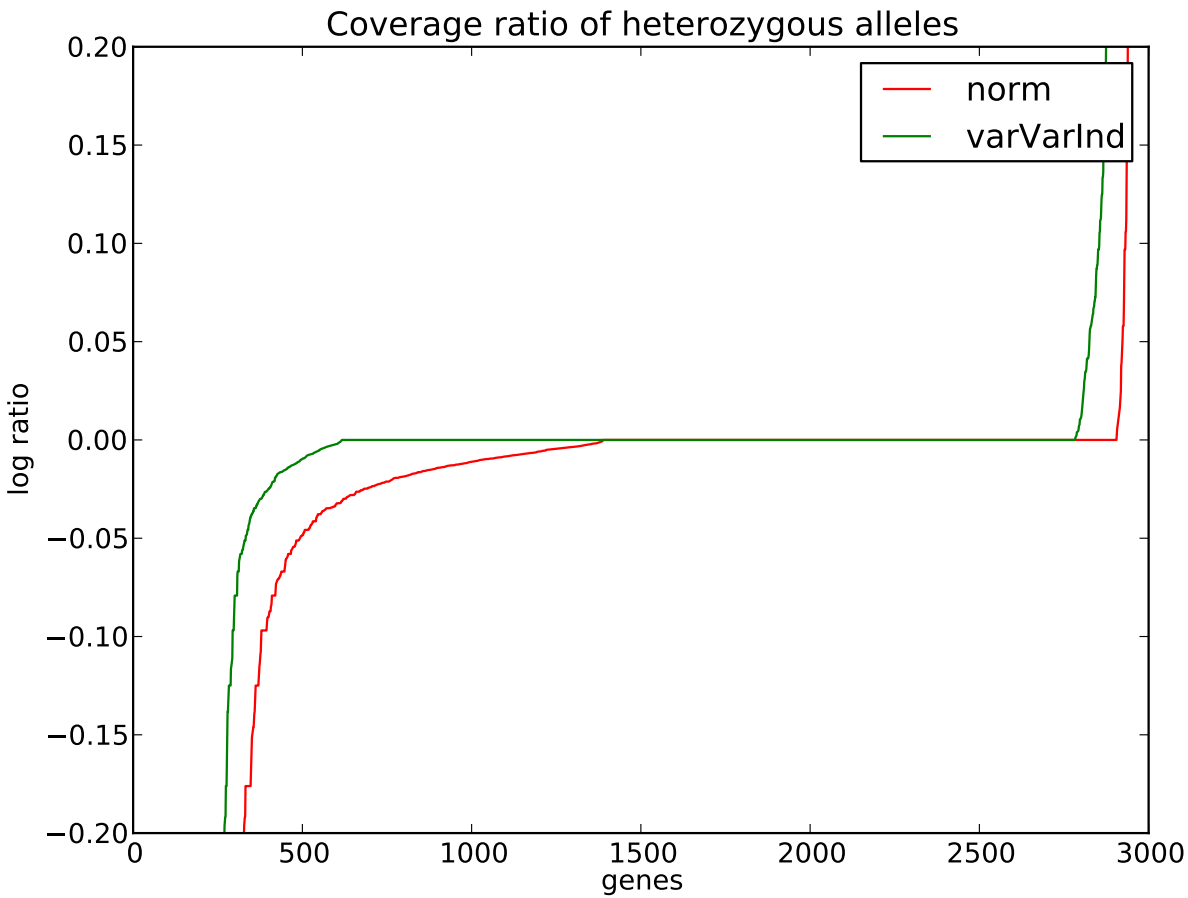
- Use FluxSimulator to generate artificial reads (10M from 5K genes)
- Duplicate and mutate read set
- Re-align to original genome
- Expectation: perfect allelic balance!



# Allelic Imbalance

## Recovery of perfect allelic balance in artificial set

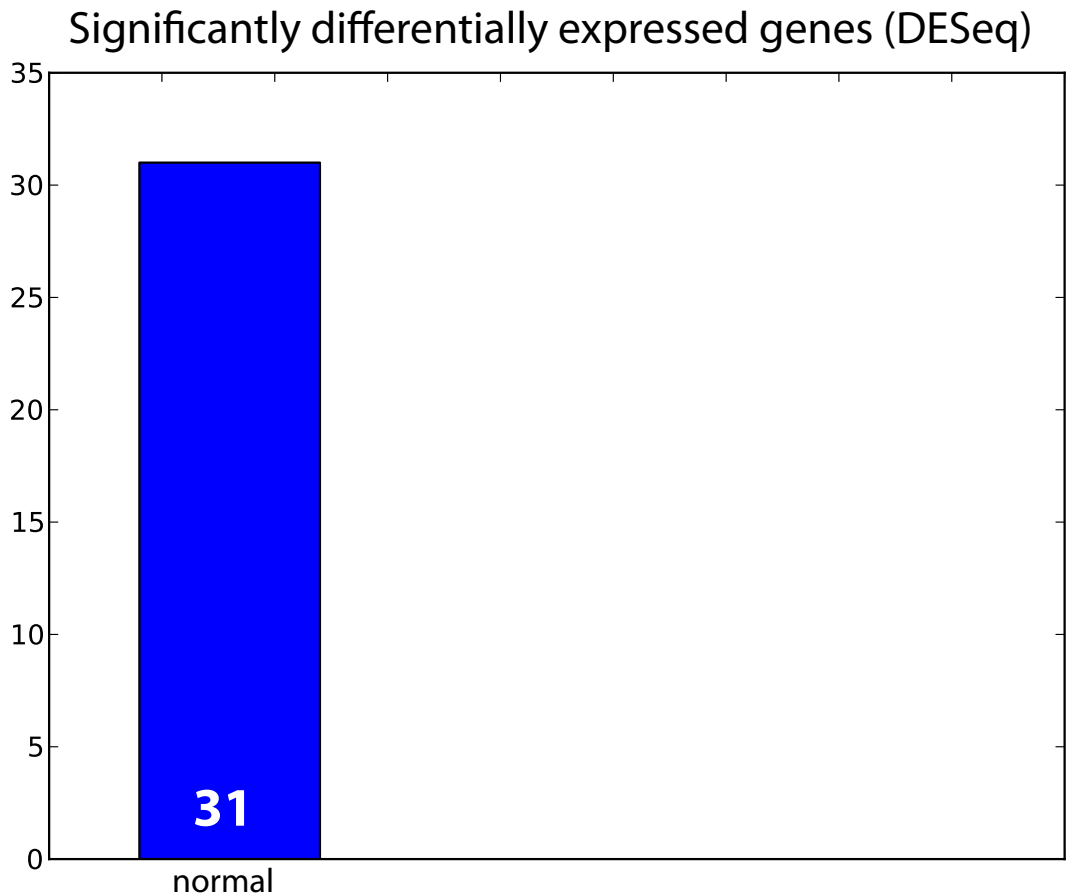
- heterozygous alleles are expected to be balanced



# Influences on Gene Expression Testing

## Measure differential gene expression with DESeq

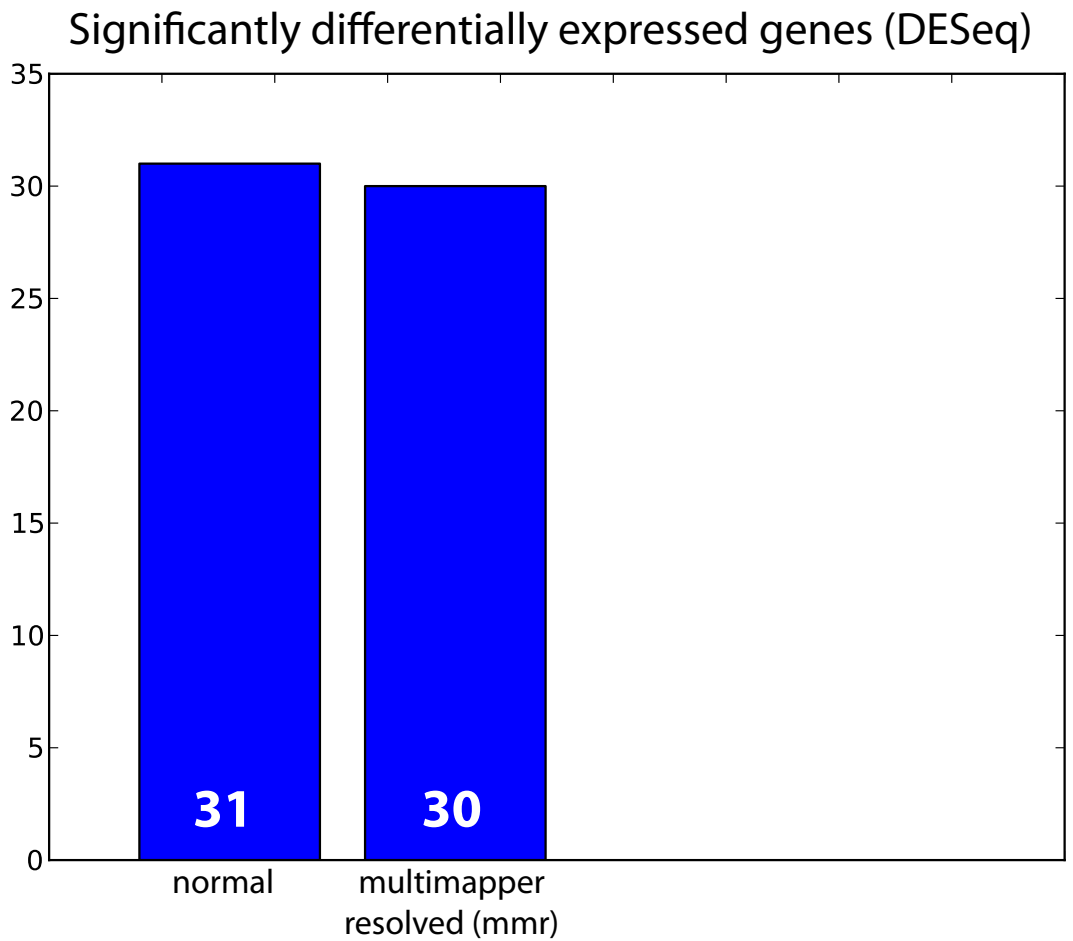
- heterozygous to reference vs. homozygous to reference



# Influences on Gene Expression Testing

## Measure differential gene expression with DESeq

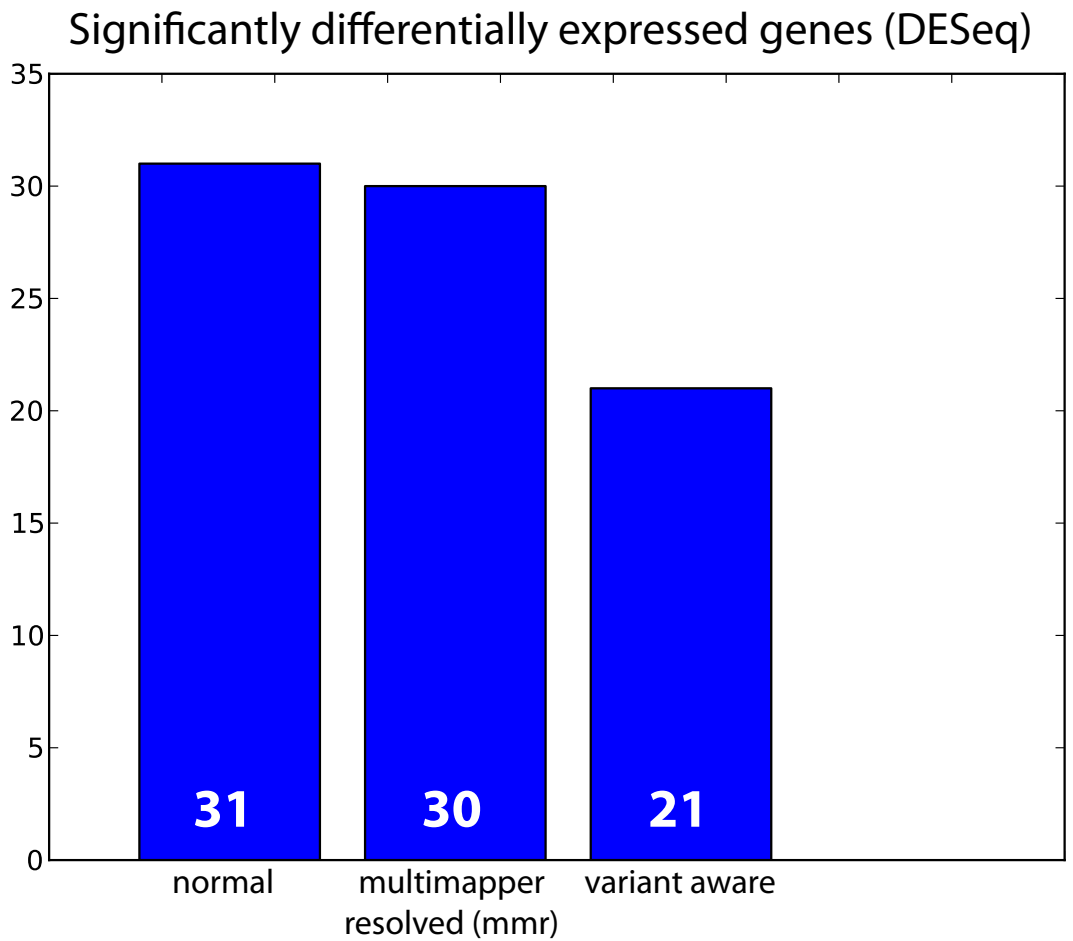
- heterozygous to reference vs. homozygous to reference



# Influences on Gene Expression Testing

## Measure differential gene expression with DESeq

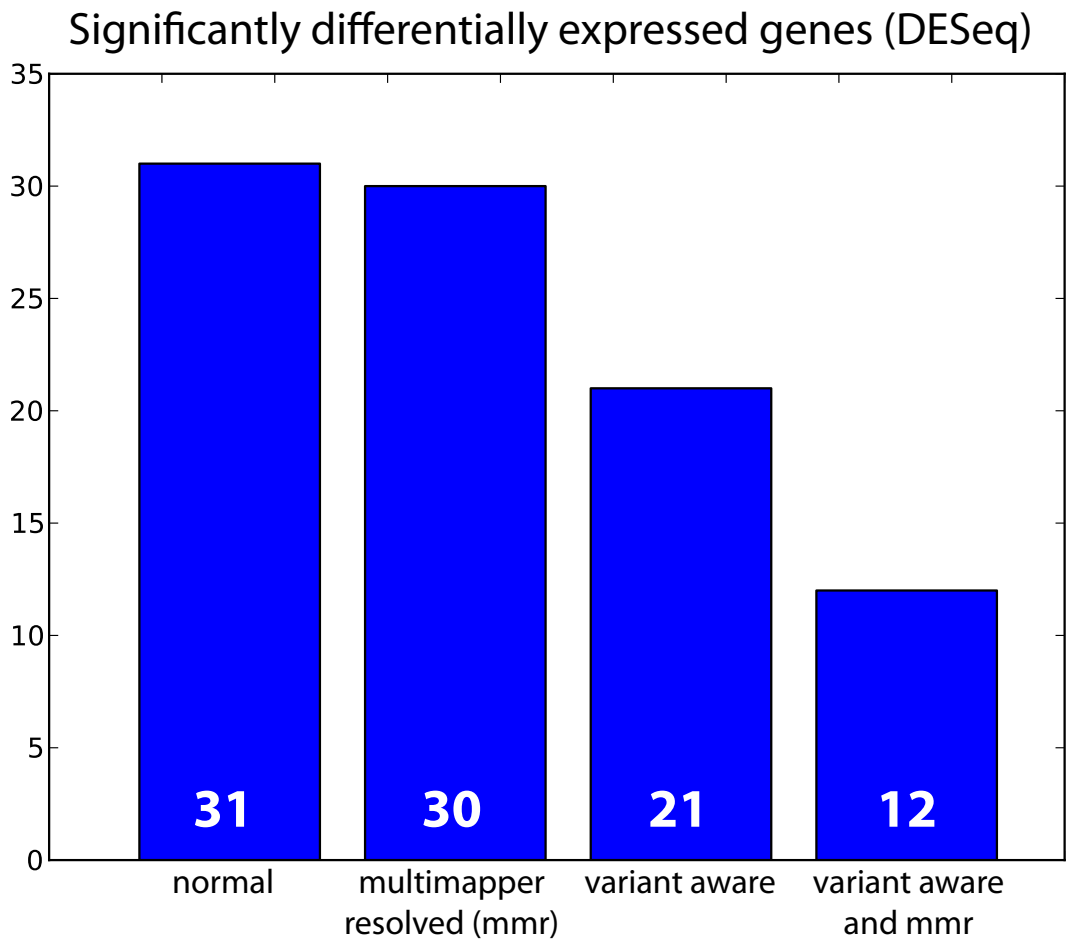
- heterozygous to reference vs. homozygous to reference



# Influences on Gene Expression Testing

## Measure differential gene expression with DESeq

- heterozygous to reference vs. homozygous to reference



# Versatile Set of Additional Features

- Fully parallelized (used tested on 64 core machines)
- Various read trimming: 3' end and polyA-tail
- Handles strand-specific reads
- Built-in filtering
- Allows to report arbitrary many multi-mappers
- Supports non-canonical splice sites
- Can align over several introns
- Junction remapping
- Multimapper resolution
- ...

# Versatile Set of Additional Features

- Fully parallelized (used tested on 64 core machines)
- Various read trimming: 3' end and polyA-tail
- Handles strand-specific reads
- Built-in filtering
- Allows to report arbitrary many multi-mappers
- Supports non-canonical splice sites
- Can align over several introns
- Junction remapping
- Multimapper resolution
- ...

# Availability

- Command-line
  `http://github.org/ratschlab/palmapper`
  `http://ftp.raetschlab.org/software/palmapper/`

- Galaxy web-interface
  `http://galaxy.cbio.mskcc.org/`

- Tutorial paper: G. Jean et al., Current Protocols in Bioinformatics, 2010.
- Original paper: A. Kahles & G. Jean et al., forthcoming.

# Acknowledgements

Gunnar Rätsch

*Supervision*

Geraldine Jean

*Variation-aware DP*

David Kuo

*Variant Parsing*

Vipin T. Sreedharan

*Galaxy Integration*

Funding by DFG & MSKCC.

Thank you for your attention.