Genomic resources

for non-model organisms

Genomic resources

- Whole genome sequencing
  - reference genome sequence
    - comparisons across species
    - identify signatures of natural selection
  - population-level resequencing
    - explore variation within species
    - identify signatures of natural selection
- Transcriptome assembly
  - reference sequences
    - comparisons across species
    - gene annotation
  - gene expression studies

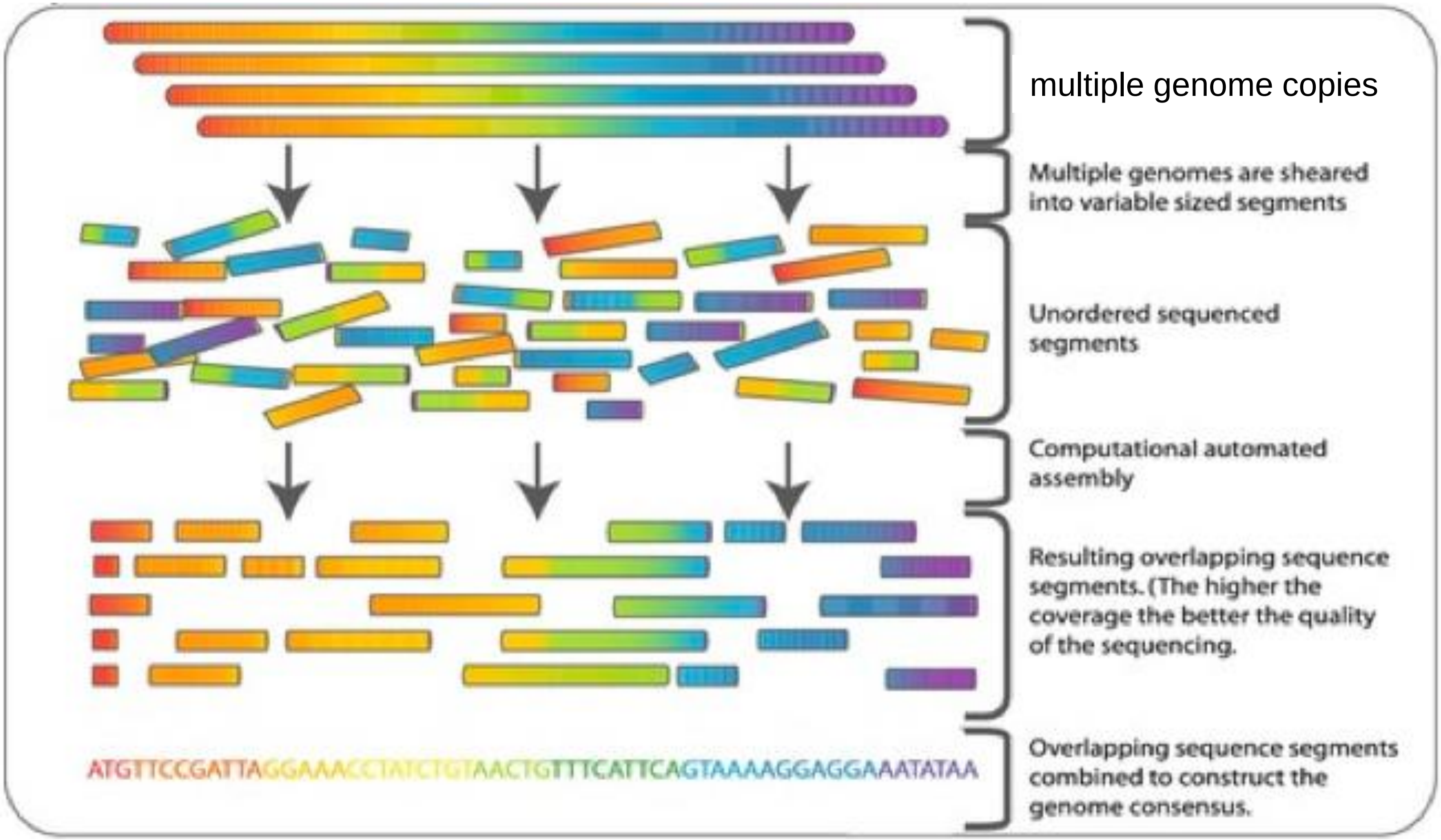
Genomic resources

- Reduced-representation sequencing (GBS)
  - compare DNA sequence variation within & between populations
    - identifying population structure and reconstructing population demographic history
  - gene mapping
    - identify genetic loci associated with traits of interest
  - forensics
    - individual identification
    - parentage tests
    - identification of optimal breeding pairs
- SNP arrays
  - same uses as above.

Whole genome sequencing (WGS)

- Sequencing technology changes constantly.
  - more reads
  - longer reads
  - lower error rates
  - cheaper.
- Popular current technology for WGS
  - Hi-C
    - Dovetail
  - Joins (ligates) pieces of DNA that are in contact within a chromosome, but that are more distant in terms of DNA sequence. Enables anchoring of DNA segments into a scaffold.

Whole genome sequencing and assembly



multiple genome copies

Multiple genomes are sheared into variable sized segments

Unordered sequenced segments

Computational automated assembly

Resulting overlapping sequence segments. (The higher the coverage the better the quality of the sequencing.

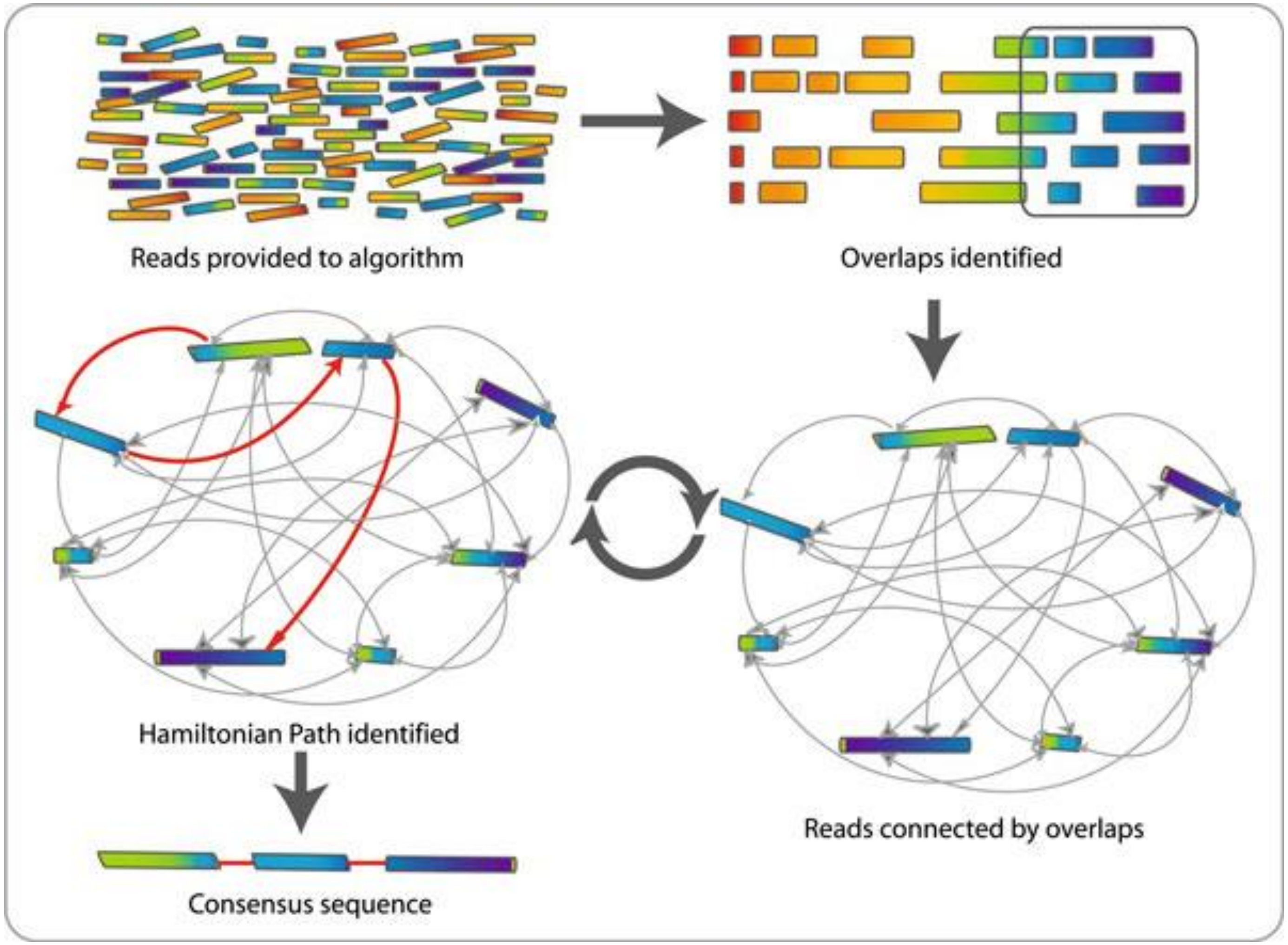
Overlapping sequence segments combined to construct the genome consensus.

doi: 10.1007/s12575-009-9004-1

Whole genome sequencing and assembly

- Tissue is collected, and DNA extracted
  - (many, many copies of the genome are represented in the sample)
- DNA is fragmented.
- Fragments are sequenced: sequence reads
- Reads with overlapping sequences are identified
  - longer sequences are assembled based on overlapping reads.

Read assembly …


Reads provided to algorithm
Overlaps identified
Reads connected by overlaps
Hamiltonian Path identified
Consensus sequence

Short reads

- Illumina
- Massive throughput
- Low error rate
- Read length ≤300nt

- Many short-read assemblers exist.
  - generate contigs that can be fairly long, but not entire chromosomes (eukaryotic)

Paired ends; mate pairs

- Longer molecules; only the ends are sequenced.
- Useful for orienting and joining contigs.

Long reads

- Single-molecule sequencing

- PacBio SMRT sequencing
  - median read length 50,000nt, some reads >175,000 nt.
  - low error rate (<1%).

- Oxford Nanopore
  - average read length 6,000-15,00nt, max read length close to 2,000,000nt.
  - fast, portable, and relatively inexpensive.
  - high error rate (~10%).

Hybrid assemblies

- Short reads + long reads.
- Short reads to generate contigs
- Long reads to join contigs.

Hybrid assemblies

- Dovetail
  - Chicago + Dovetail

dovetailgenomics.com/ga_tech_overview

Hurdles to assembly

A number of factors increase the difficulty of creating a correct assembly

- High heterozygosity
- Repetitive regions
- Genome duplications
- Polyploidy

- If possible, use an accession that is diploid and inbred (low heterozygosity) to create the reference
  - Can then use this to aid genomics/transcriptomics of more complex accessions/species

Non-model organisms: issues

- Difficulties acquiring samples
- Small sample sizes
- DNA/RNA quality from “non-standard” samples
    - small quantities of tissues/blood
    - feces
    - remains

Genotyping by sequencing (GBS)

- The idea: sequence your samples’ genomes and compare sequence variation across samples
  - identify variable sites
  - call genotypes at these sites
- Coverage & accuracy vs. cost
  - the deeper the coverage, the more reliable the genotype calls, and the higher the per sample cost.
- The higher the heterozygosity, the lower the accuracy
  - need good coverage to reliably distinguish heterozygotes from sequencing error.

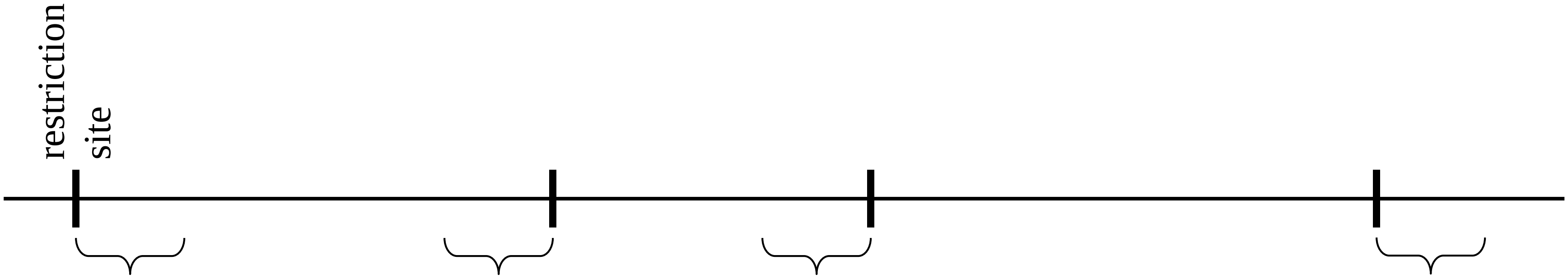
Genotyping by sequencing (GBS)

- Full genome sequencing
  - may be reasonable if the genome is very small or a good reference genome is available.
  - (currently) prohibitively expensive if the genome size is moderate or large and no reference is available.
- Instead of sequencing the entire genome, focus on particular regions (reduced representation libraries)
  - e.g. exome
    - exon capture
    - mRNA
  - or random sections of the genome
    - e.g. RAD-tag sequencing

Genotyping by sequencing (GBS)

RAD-tag sequencing

- Focus on high-depth sequencing of a small fraction of the genome:
  - short sections of DNA directly adjacent to specific restriction enzyme recognition sites
- Restriction-site Associated DNA (RAD)



restriction site

only regions next to restriction sites are sequenced

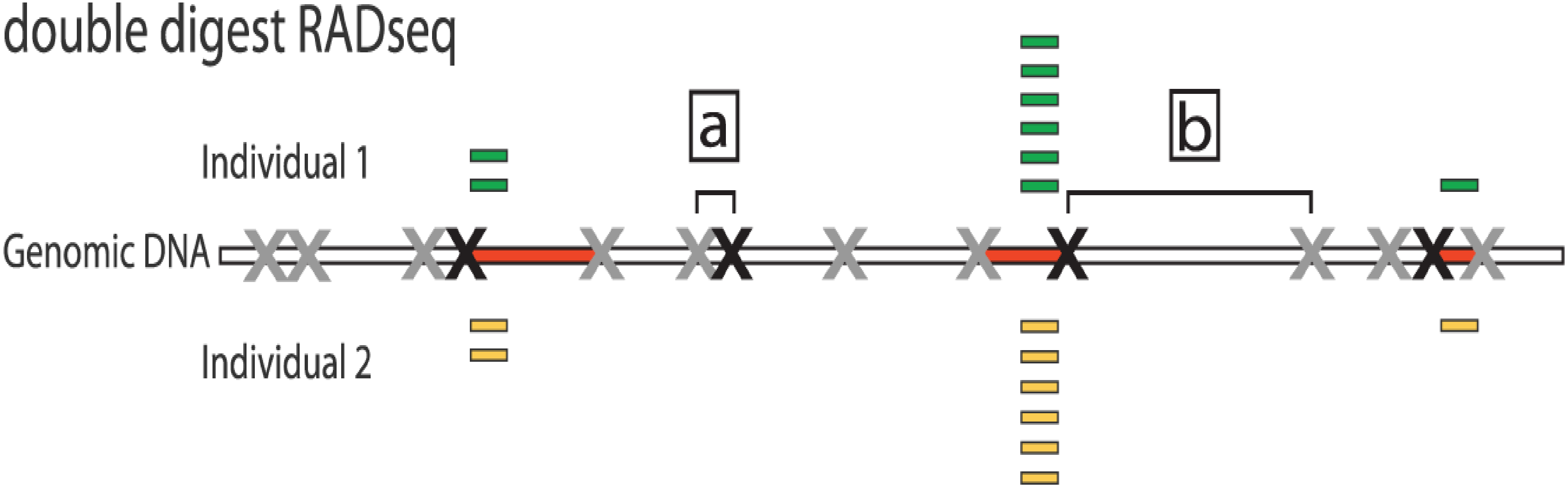
Baird et al, Plos ONE (2009) 3(10): e3376

GBS: RADseq

- Extract genomic DNA, cut with restriction enzymes:
  - one common, one rare.
- Size select fragments
  - one end containing rare restriction site, one with common restriction site.
- Ligate adapters to ends
  - (includes Illumina sequencing primer)
- Amplify fragments that contain adapter bound to restriction site
- Sequence from end of fragment with the rare restriction site.

Peterson et al, PLoS ONE (2012) 7(5): e37135

GBS: RADseq



doi.org/10.1371/journal.pone.0037135.g002

RADseq downstream analyses

- If a reference genome sequence is available, reads are aligned to the reference.
- If no reference genome is available, assembly-like algorithms are used.
  - e.g. Stacks (creskolab.uoregon.edu/stacks), rtd (github.com/brantp/rtd)
  - These take advantage of the fact that only a small portion of the genome has been sequenced (at high coverage)
  - Sequencing is expected to start at the same nucleotide location for each region of the genome that was targeted.
    - (reads largely overlapping, not tiled)
  - Autopolyploids (no reference), feasibility unclear.

GBS: Skim sequencing

- Generally relies on having a reference genome
  - possibly also already known marker sites.
- Sequence genomic DNA
  - low coverage (fewer reads)
- Align reads to genome
- Marker/genotype calling software

GBS: marker ID & geno calls

- Sites where a sufficient number of aligned or assembled reads contain sequence differences are determined to be polymorphic.
- The proportion of reads containing each allelic sequence determines genotype status:
    - 100% (or close to) indicates a homozygote
    - proportions somewhere around 50% one type/50% the other indicates a heterozygote in a diploid species.
    - For polyploids, various ratios are possible.
        - some methods exist (e.g. Garcia, et al., 2013, Sci. Rep, 3:3399)
        - software underdeveloped
        - pipelines described (e.g. Saintenac, et.al., 2013, G3 3:1105-1114)

Transcriptomes: RNA-Seq

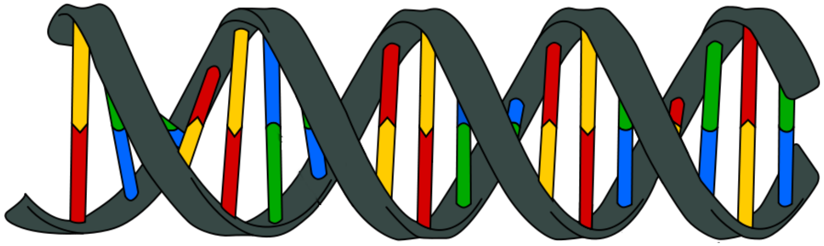
- RNA-Seq
    - sequencing of transcripts
- Gene expression studies
    - compare expression across conditions
        - time, developmental stages, genotypes
- Compare transcriptome sequences across species
- Identify sequence variation within populations.

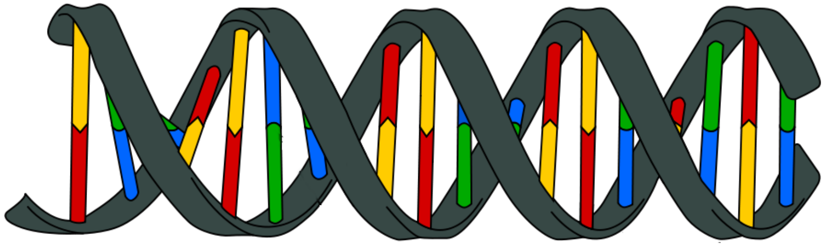
Gene expression

- Measured through transcript (mRNA) abundance

Condition A

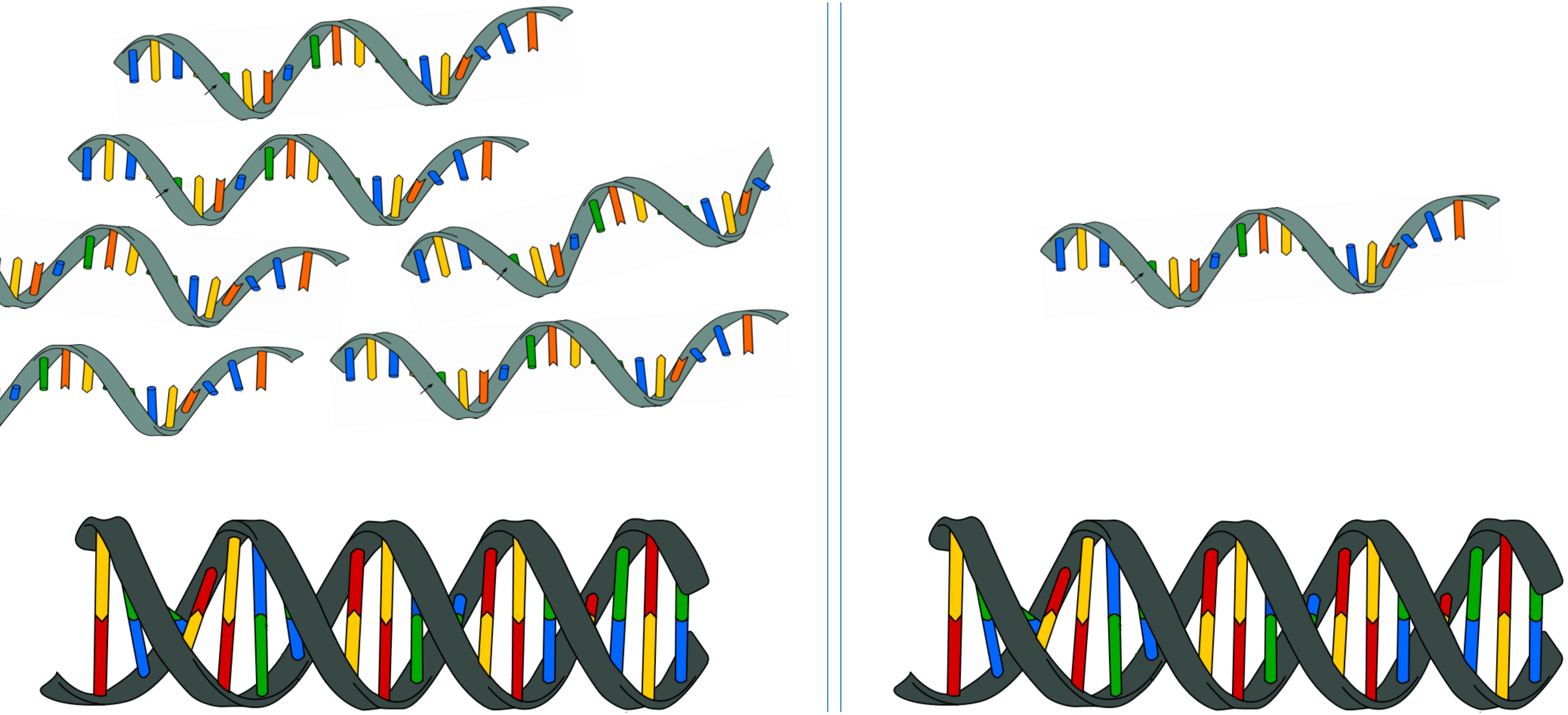
Condition B





Gene expression

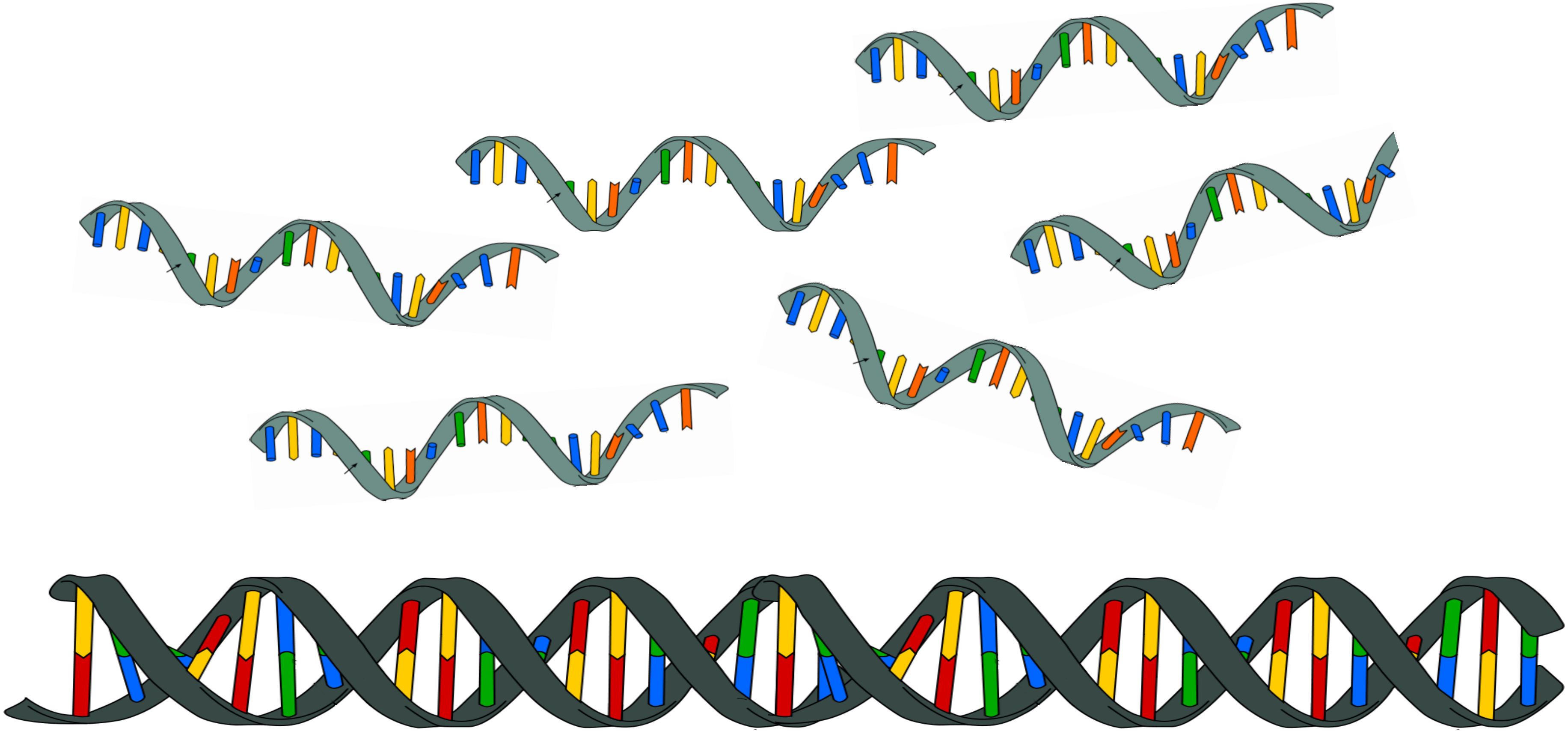
- Measured through transcript (mRNA) abundance



Gene expression: RNA-Seq

- Collect biological sample
- extract mRNA
- ultra-high throughput sequencing
  - each mRNA molecule that was sampled for sequencing produces a sequence read
- if a gene was highly expressed in the sample
  - transcript abundance is high
- many sequence reads will be generated for that gene (relative to other genes)

Sequence reads



Sequence reads

GTTAAGGCTGCCATCAAGGACAGGGTTGTCAATGTTGCTCAAGTTACCAGCAACACACTCGCTTT

CAACAAGAGAAAACAAGGTGCAAGTATTGCCTTGGAACTGGTTACTTGGCTTGCGCTCGGTGTTC

CGGGAAACCAAATCAAGAAGCAGGCAATCCTTAGGATTGCTTTTCGTGGGTAGAGCGAGGGTTT

ATTTTTCAGTCTTCTCTCGTGGCATTTATTGTCGGTTGGTTTTCTATATATTGCTCGTGCAACTC

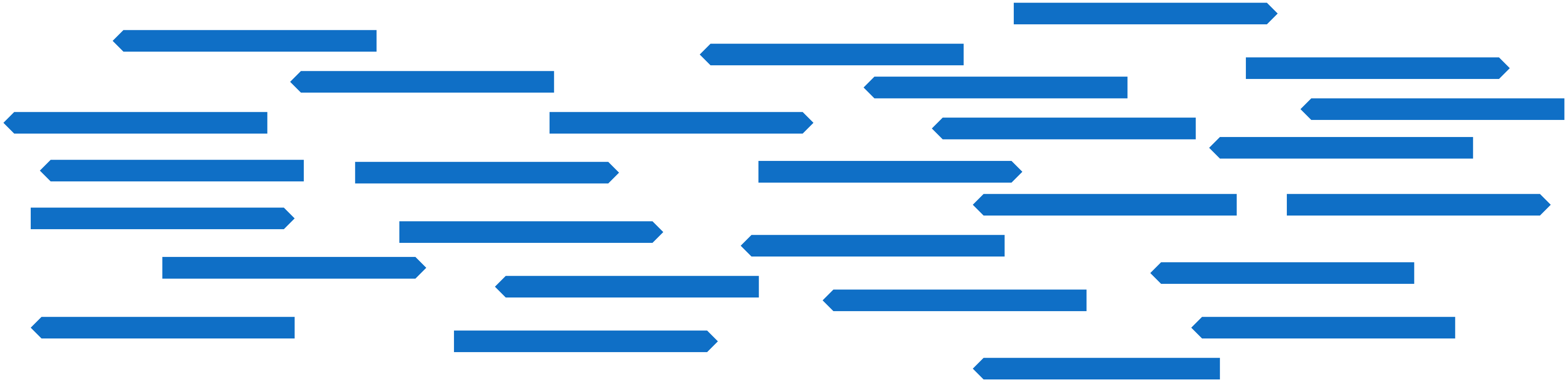
CGTCCCTACCATATCTCATCATCATTATCAATAATATAAGAAACATAATTATCATAATAGAGGAA

CTCTTGCCGGCATTGTGGGCAAAGAGAGAATTGTTGTGTCCACTTCTTGCTCACTTCTTCACACT

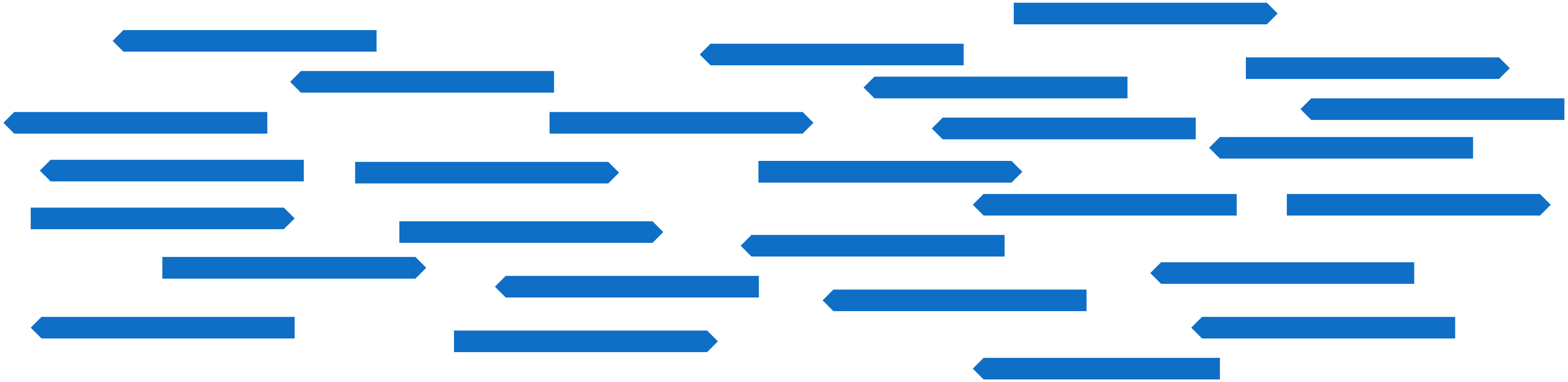
TGTCATAATAACACTCTCTGCTGGTAGAGGTGCAGAATGCTGTAACATACCATCCCCTTCTTTTA

AAAATATATTTCTGGGATCAATTGACAAAGGATGATATCAAAGTGTACGGATATGTTTCTGAGA

Millions of short sequence reads



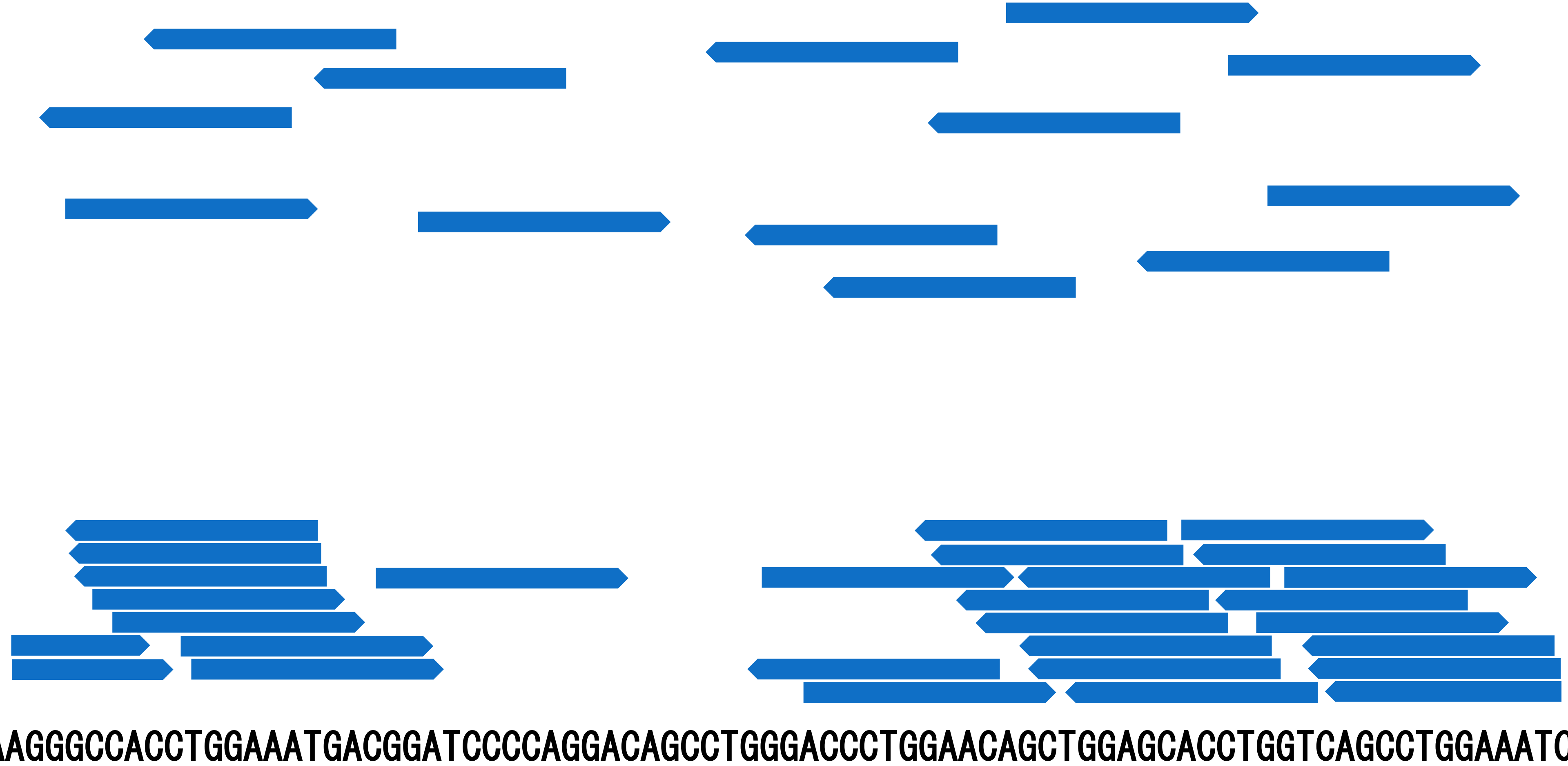
Millions of short sequence reads



AAGGGCCACCTGGAAATGACGGATCCCCAGGACAGCCTGGGACCCTGGAACAGCTGGAGCACCTGGTCAGCCTGGAAATC

genome sequence

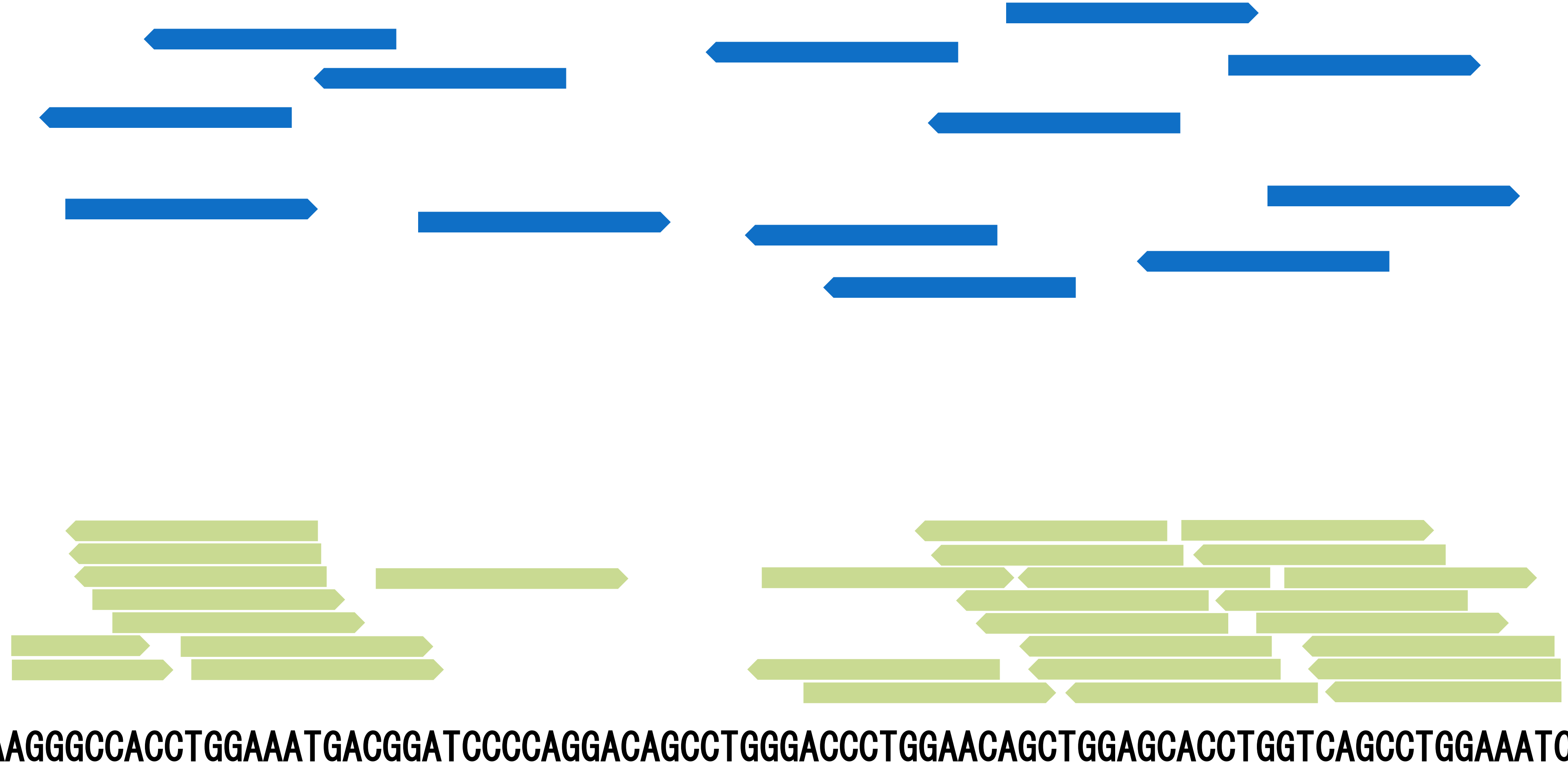
Align short reads to genome



AAGGGCCACCTGGAAATGACGGATCCCCAGGACAGCCTGGGACCCTGGAACAGCTGGAGCACCTGGTCAGCCTGGAAATC

genome sequence

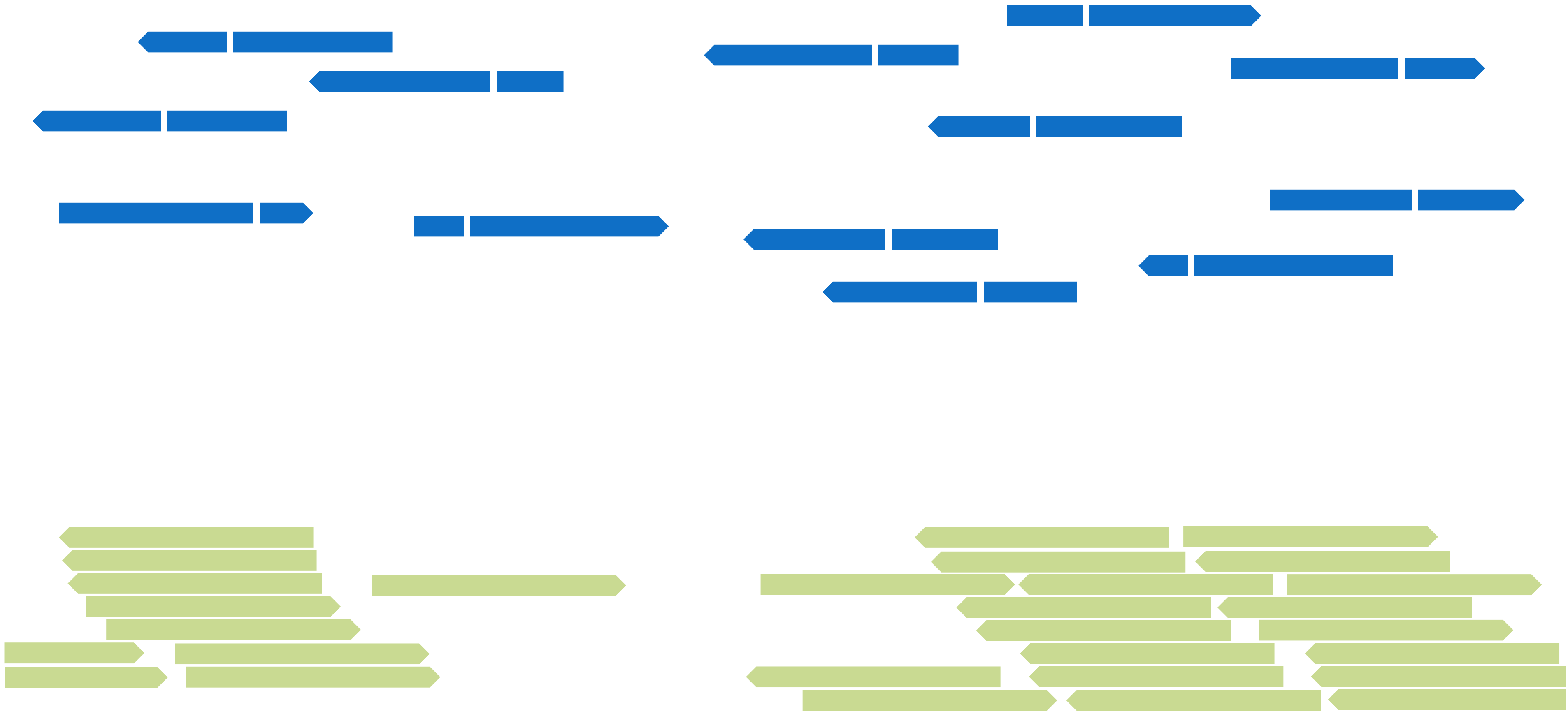
Reads that don’t align in first pass …



AGGGCCACCTGGAAATGACGGATCCCCAGGACAGCCTGGGACCCTGGAACAGCTGGAGCACCTGGTCAGCCTGGAAATC

genome sequence

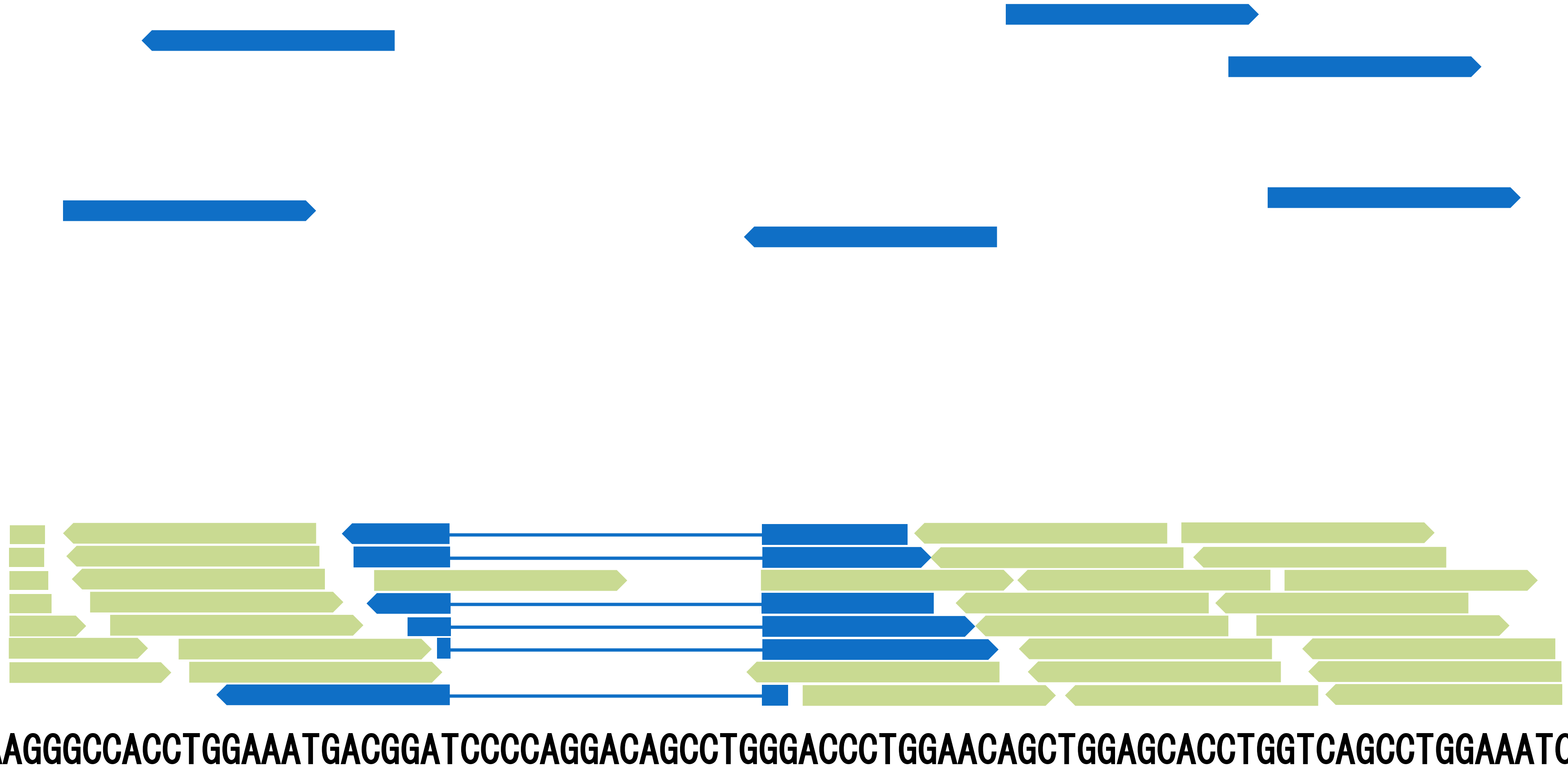
Break into pieces



AAGGGCCACCTGGAAATGACGGATCCCCAGGACAGCCTGGGACCCTGGAACAGCTGGAGCACCTGGTCAGCCTGGAAATC

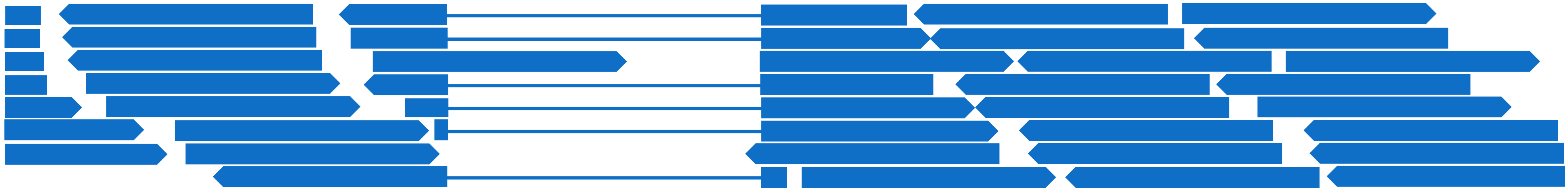
genome sequence

Align allowing for gaps: introns


AAGGGCCACCTGGAAATGACGGATCCCCAGGACAGCCTGGGACCCTGGAACAGCTGGAGCACCTGGTCAGCCTGGAAATC

genome sequence

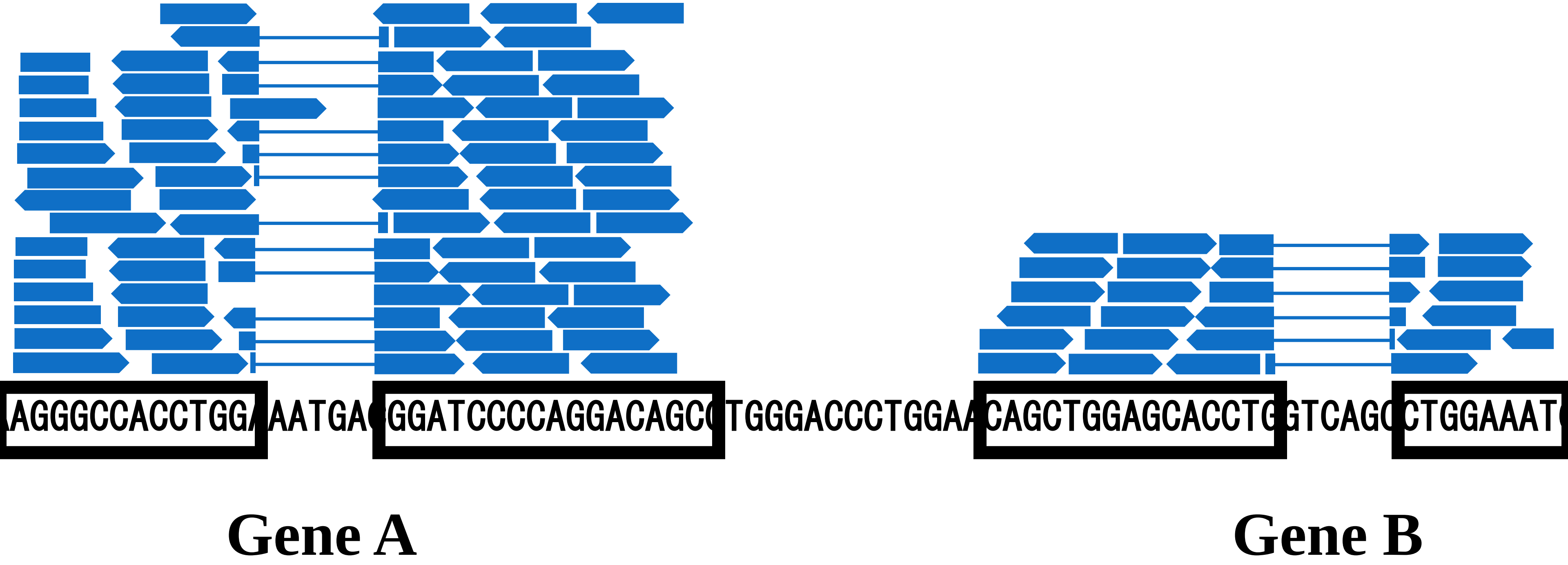
Use alignments to determine which genes contributed which sequenced transcripts



AAGGGCCACCTGGAAATGACGGATCCCCAGGACAGCCTGGGACCCTGGAACAGCTGGAGCACCTGGTCAGCCTGGAAATC

genome sequence

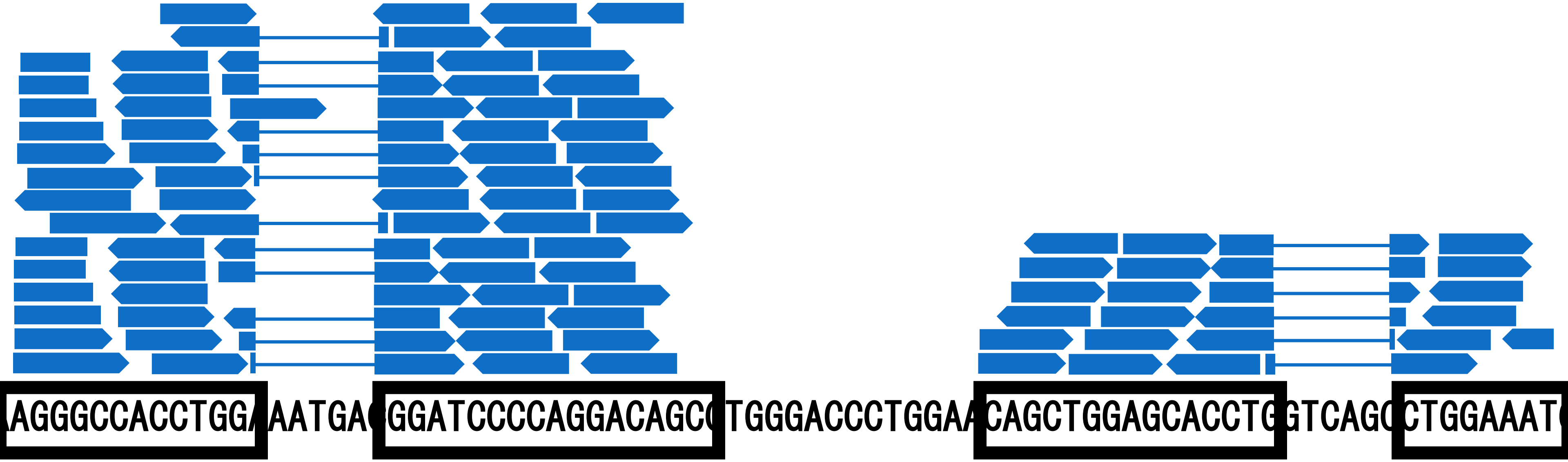
Use alignments to determine which genes contributed which sequenced transcripts


AGGGCCACCTGGAAATGACGGATCCCCAGGACAGCCTGGGACCCTGGAACAGCTGGAGCACCTGGGTCAGCCTGGAAATC
Gene A
Gene B

And for quantification of gene expression (counts of reads per gene)

Gene A: 97

Gene B: 32



AGGGCCACCTGGAAATGACGGATCCCCAGGACAGCCTGGGACCCTGGAACAGCTGGAGCACCTGGGTCAGGCTGGAAATG

Gene A

Gene B

Reference sequences

- What happens if you don’t have a reference genome available
- And you don’t have the resources to generate one

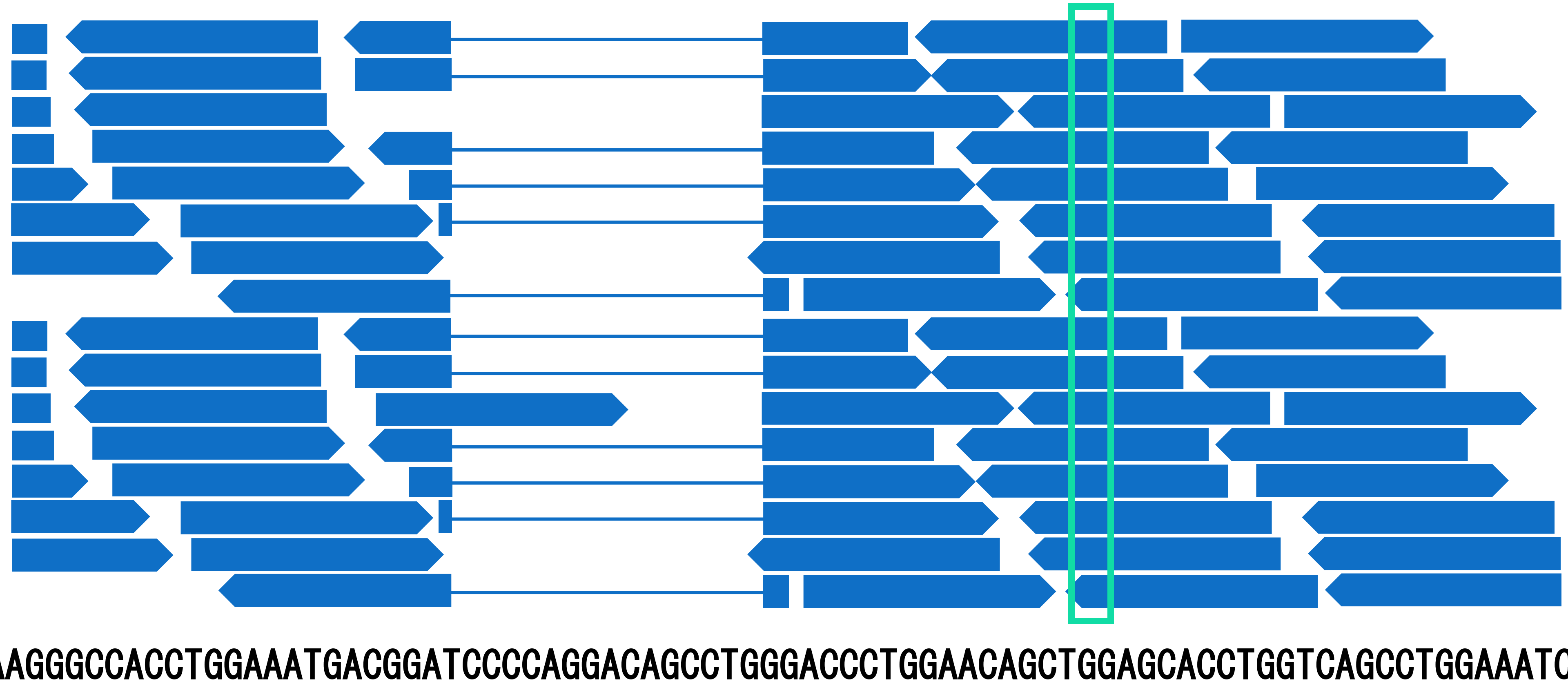
Transcriptome assembly

- Use RNA-Seq reads to create a transcriptome reference
    - assembly process is similar to the process for whole genome assembly
        - (different software)
- End product: predicted sequences of transcribed regions
    - exons only
    - different entries for splice variants
- Use this as a reference to compute transcript abundance (quantification)

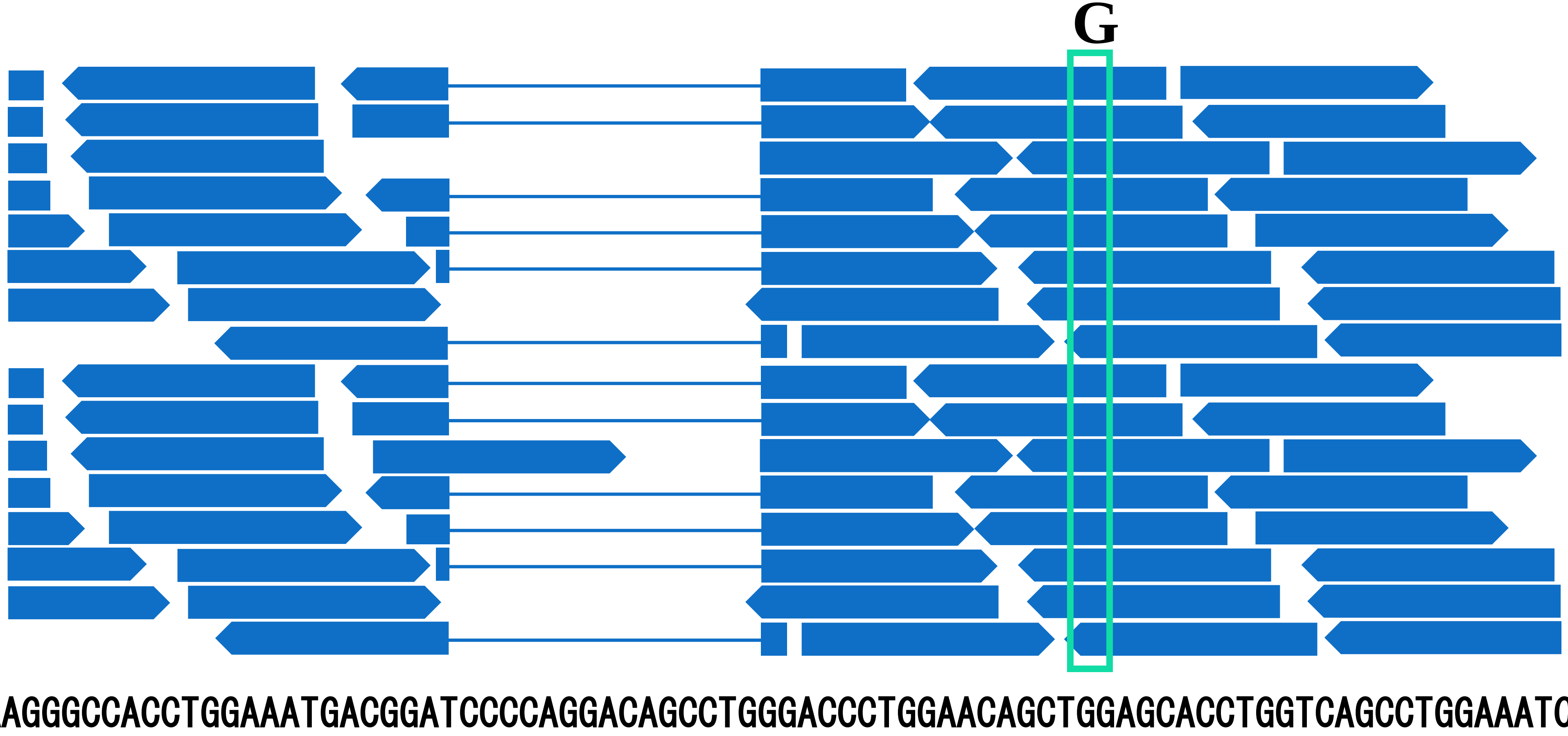
RNA-Seq data

also provides the ability to locate sequence variation across individuals

Identifying sequence polymorphisms


AGGGCCACCTGGAAATGACGGATCCCCAGGACAGCCTGGGACCCTGGAACAGCTGGAGCACCTGGTCAGCCTGGAAATC

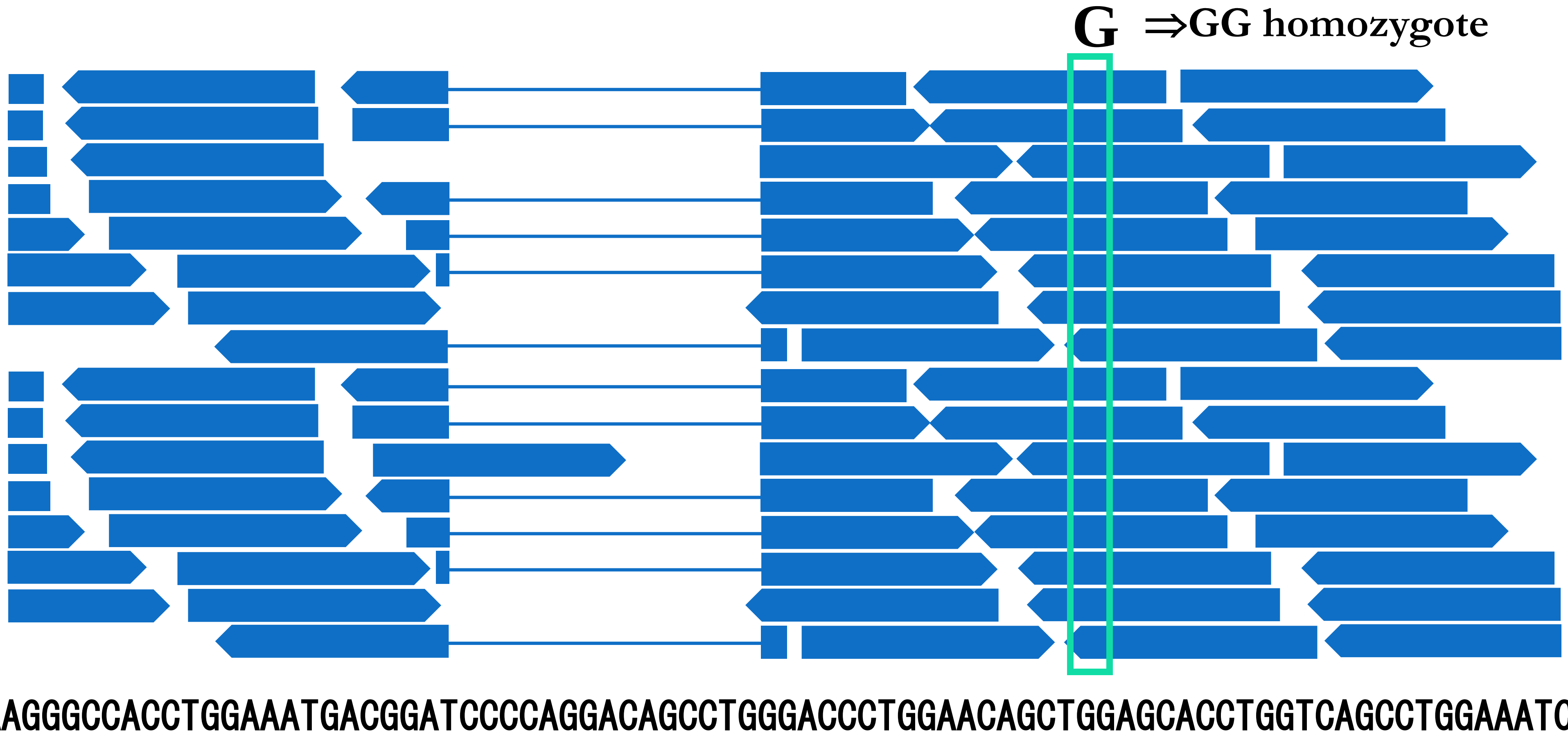
Identifying sequence polymorphisms



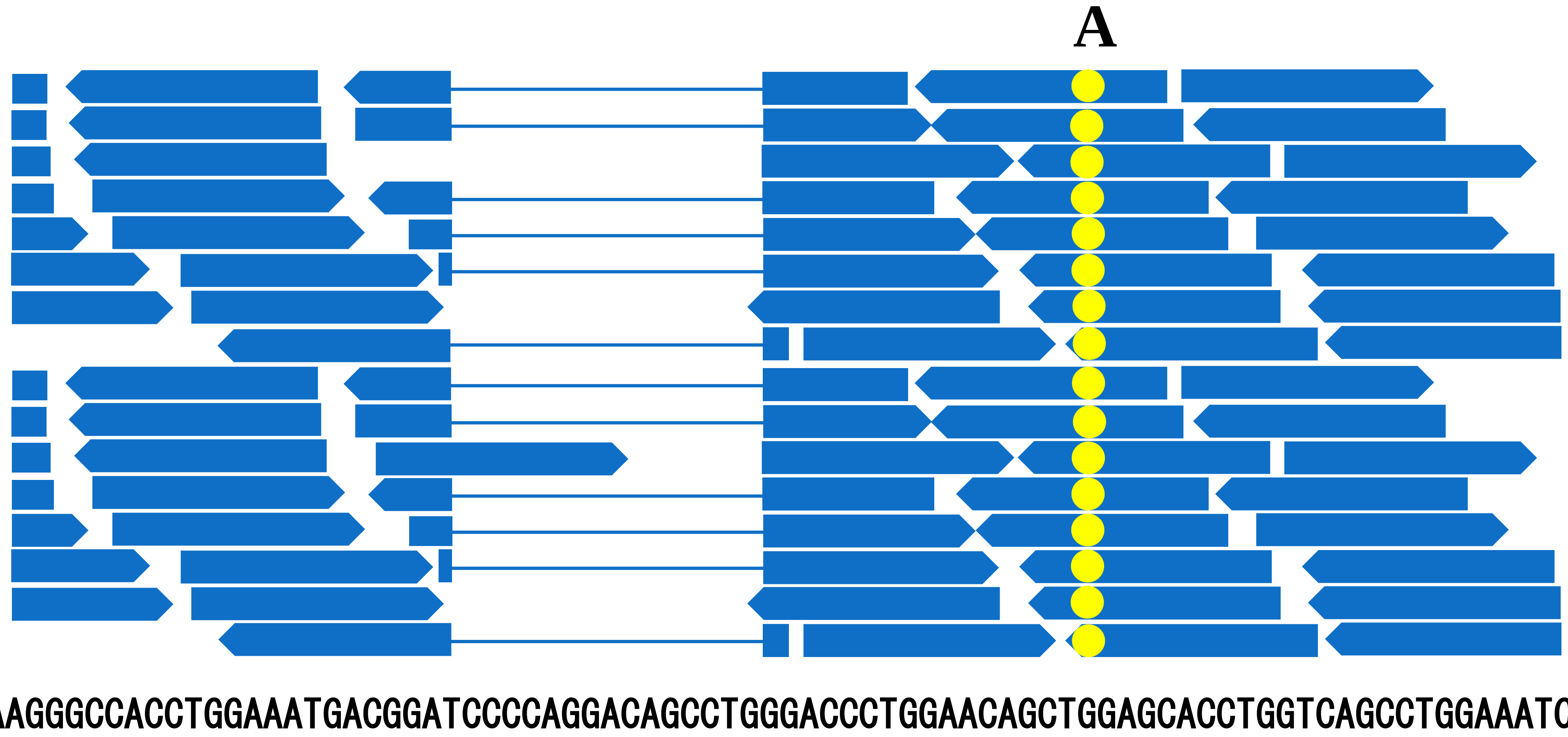
G

AAGGGCCACCTGGAAATGACGGATCCCCAGGACAGCCTGGGACCCTGGAACAGCTGGAGCACCTGGTCAGCCTGGAAATC

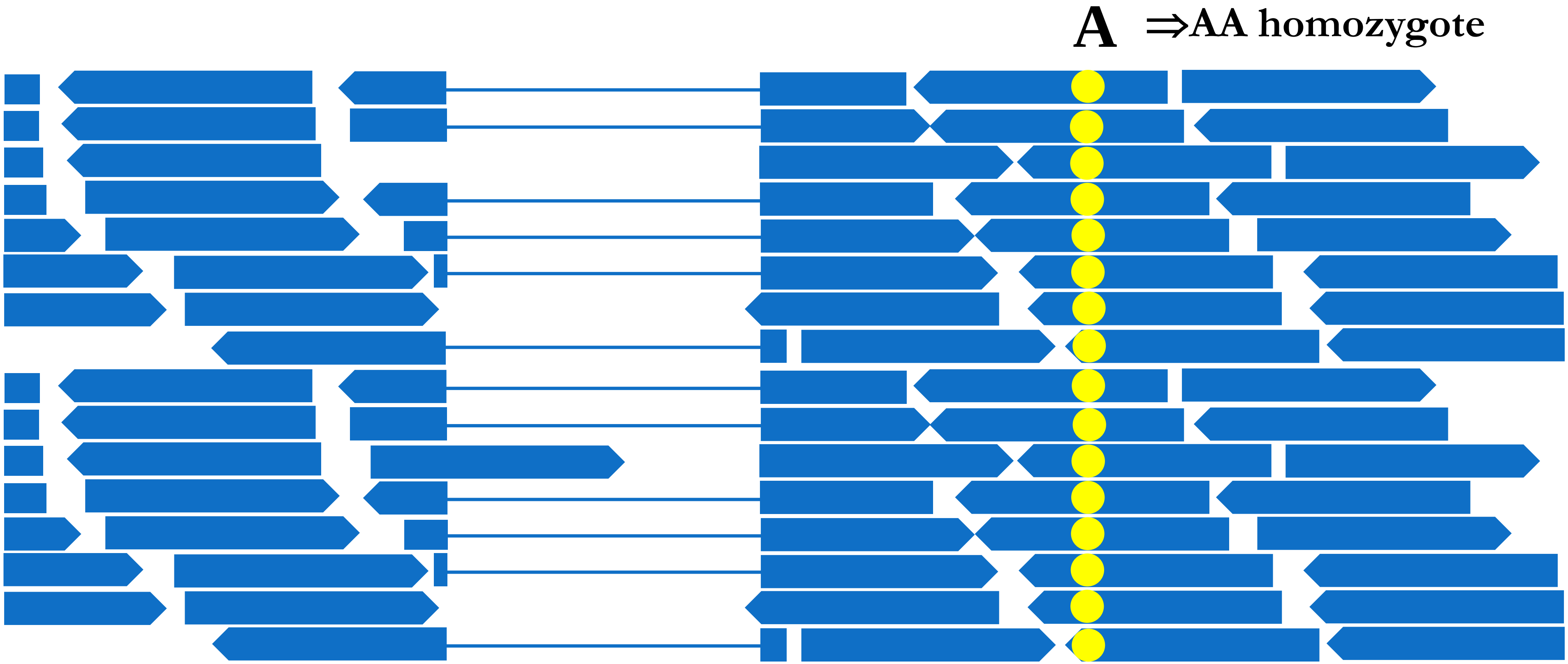
Identifying sequence polymorphisms


G
⇒GG homozygote
AAGGGCCACCTGGAAATGACGGATCCCCAGGACAGCCTGGGACCCTGGAACAGCTGGAGCACCTGGTCAGCCTGGAAATC

Identifying sequence polymorphisms


A
AAGGGCCACCTGGAAATGACGGATCCCCAGGACAGCCTGGGACCCTGGAACAGCTGGAGCACCTGGTCAGCCTGGAAATC

Identifying sequence polymorphisms

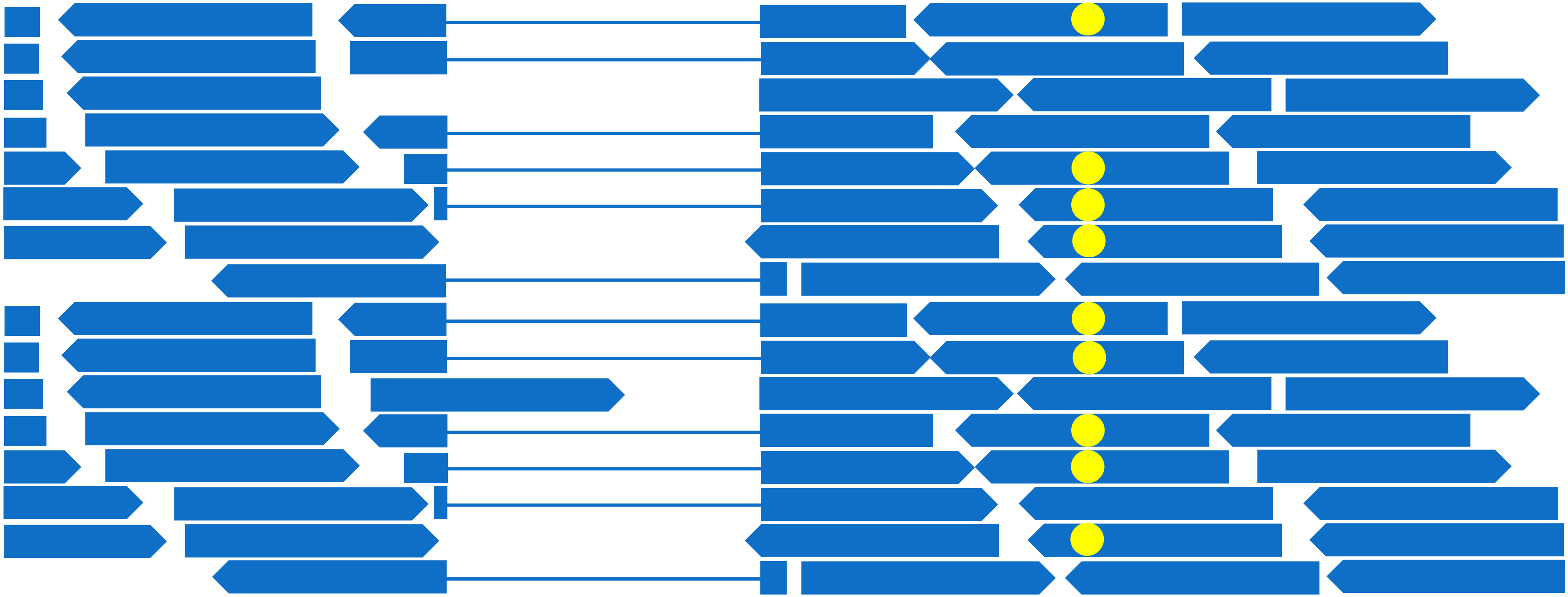


A ⇒AA homozygote

AGGGCCACCTGGAAATGACGGATCCCCAGGACAGCCTGGGACCCTGGAACAGCTGGAGCACCTGGTCAGCCTGGAAATC

Heterozygote (~50% of each allele)

⇒AG heterozygote



AAGGGCCACCTGGAAATGACGGATCCCCAGGACAGCCTGGGACCCTGGAACAGCTGGAGCACCTGGTCAGCCTGGAAATC

AS YOU CAN CLEARLY SEE IN SLIDE 397...
scottadams@aol.com
www.dilbert.com
GAAAAH!
Z
Z
© 2000 United Feature Syndicate, Inc.
8/16/00
"POWERPOINT" POISONING.

http://dilbert.com/strip/2000-08-16