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

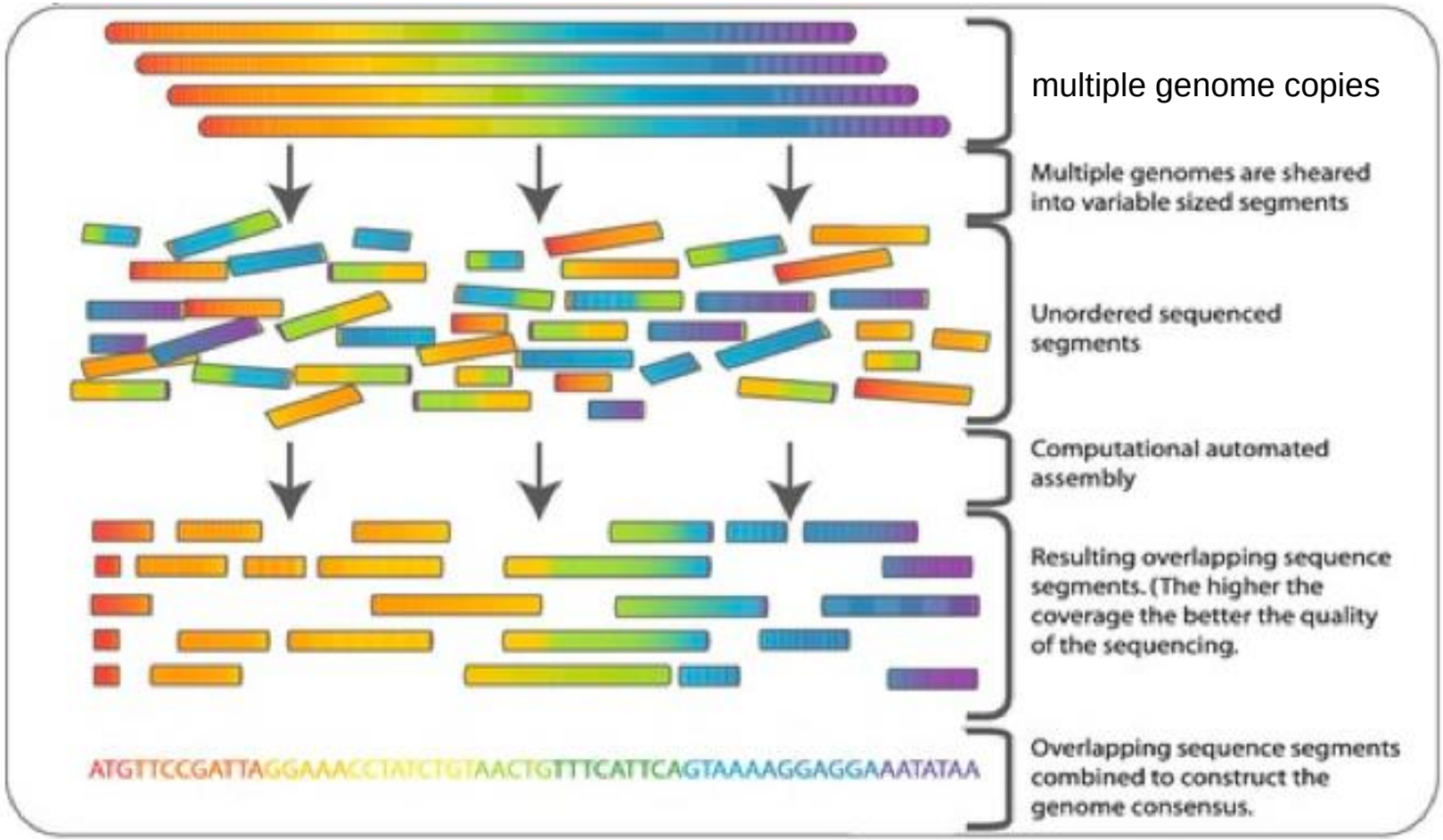
Genomic resources

- 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.
- Current “it” technology for WGS …
  - Hi-C
    - Dovetail
  - Joins (ligates) pieces of DNA that are in contact within a chromosome, but that are distant in terms of DNA sequence. Enables anchoring of DNA segments into a scaffold.

Whole genome sequencing and assembly

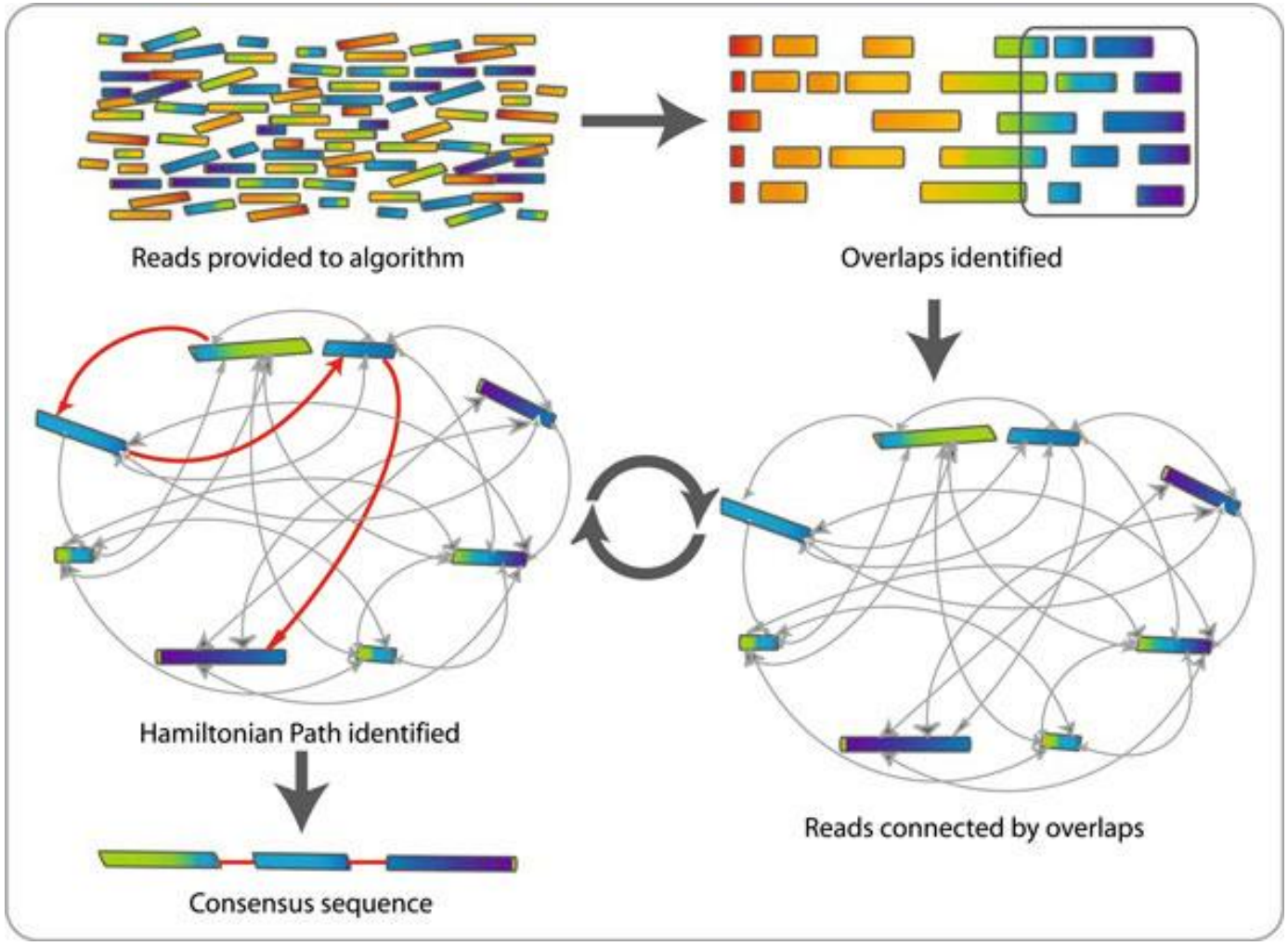


doi: 10.1007/s12575-009-9004-1

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 …



Short reads

- Illumina
- Massive throughput
- Low error rate
- Read length <200nt

- 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
  - average read length >15,000nt, some reads >100,000 nt.
  - current technology provides low error rates for shorter range of reads.

- Oxford Nanopore
  - average read length 6,000-15,00nt, max read length close to 1,000,000nt.
  - fast, portable, and relatively inexpensive.
  - high error rate (10-15%).

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

Title: Genome sequence and population declines in the critically endangered greater bamboo lemur (Prolemur simus) and implications for conservation
Source: BMC GENOMICS **Volume:** 19 **Article Number:** 445 **DOI:** 10.1186/s12864-018-4841-4 **Published:** JUN 8 2018
PubMed ID: 29884119

Title: The inference of gray whale (Eschrichtius robustus) historical population attributes from whole-genome sequences
Source: BMC EVOLUTIONARY BIOLOGY **Volume:** 18 **Article Number:** 87 **DOI:** 10.1186/s12862-018-1204-3 **Published:** JUN 7 2018
PubMed ID: 29879895

Title: Draft genome sequence of ramie, Boehmeria nivea (L.) Gaudich
Source: MOLECULAR ECOLOGY RESOURCES **Volume:** 18 **Issue:** 3 **Pages:** 639-645 **DOI:** 10.1111/1755-0998.12766 **Published:** MAY 2018
PubMed ID: 29423997

Title: The draft genome sequence of forest musk deer (Moschus berezovskii)
Source: GIGASCIENCE **Volume:** 7 **Issue:** 4 **DOI:** 10.1093/gigascience/giy038 **Published:** APR 9 2018
PubMed ID: 29635287

Title: Genome Sequence of the Freshwater Yangtze Finless Porpoise
Source: GENES **Volume:** 9 **Issue:** 4 **Article Number:** 213 **DOI:** 10.3390/genes9040213 **Published:** APR 2018
PubMed ID: 29659530

Title: Draft genome sequence of the Tibetan medicinal herb Rhodiola crenulata
Source: GIGASCIENCE **Volume:** 6 **Issue:** 6 **DOI:** 10.1093/gigascience/gix033 **Published:** MAY 5 2017
PubMed ID: 28475810

Title: The genome sequence of the wisent (Bison bonasus)
Source: GIGASCIENCE **Volume:** 6 **Issue:** 4 **DOI:** 10.1093/gigascience/gix016 **Published:** MAR 10 2017

Title: Genome sequence, population history, and pelage genetics of the endangered African wild dog (Lycaon pictus)
Source: BMC GENOMICS **Volume:** 17 **Article Number:** 1013 **DOI:** 10.1186/s12864-016-3368-9 **Published:** DEC 9 2016
PubMed ID: 27938335

Title: Development and characterization of genic SSR markers from low depth genome sequence of Clarias batrachus (magur)
Source: JOURNAL OF GENETICS **Volume:** 95 **Issue:** 3 **Pages:** 603-609 **DOI:** 10.1007/s12041-016-0672-8 **Published:** SEP 2016
PubMed ID: 27659331

Title: Whole-genome sequence analysis shows that two endemic species of North American wolf are admixtures of the coyote and gray wolf
Source: SCIENCE ADVANCES **Volume:** 2 **Issue:** 7 **Article Number:** UNSP e1501714 **DOI:** 10.1126/sciadv.1501714 **Published:** JUL 2016
PubMed ID: 29713682

Title: Genome-Wide Analysis of Simple Sequence Repeats and Efficient Development of Polymorphic SSR Markers Based on Whole Genome Re-Sequencing of Multiple Isolates of the Wheat Stripe Rust Fungus
Source: PLOS ONE **Volume:** 10 **Issue:** 6 **Article Number:** e0130362 **DOI:** 10.1371/journal.pone.0130362 **Published:** JUN 12 2015
PubMed ID: 26068192

Title: Genome Sequence of Rhizobium sp Strain CCGE510, a Symbiont Isolated from Nodules of the Endangered Wild Bean Phaseolus albescens
Source: JOURNAL OF BACTERIOLOGY **Volume:** 194 **Issue:** 22 **Pages:** 6310-6311 **DOI:** 10.1128/JB.01536-12 **Published:** NOV 2012
PubMed ID: 23105056

Title: Functional annotation from the genome sequence of the giant panda
Source: PROTEIN & CELL **Volume:** 3 **Issue:** 8 **Pages:** 602-608 **DOI:** 10.1007/s13238-012-2914-8 **Published:** AUG 2012
PubMed ID: 22865348

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.

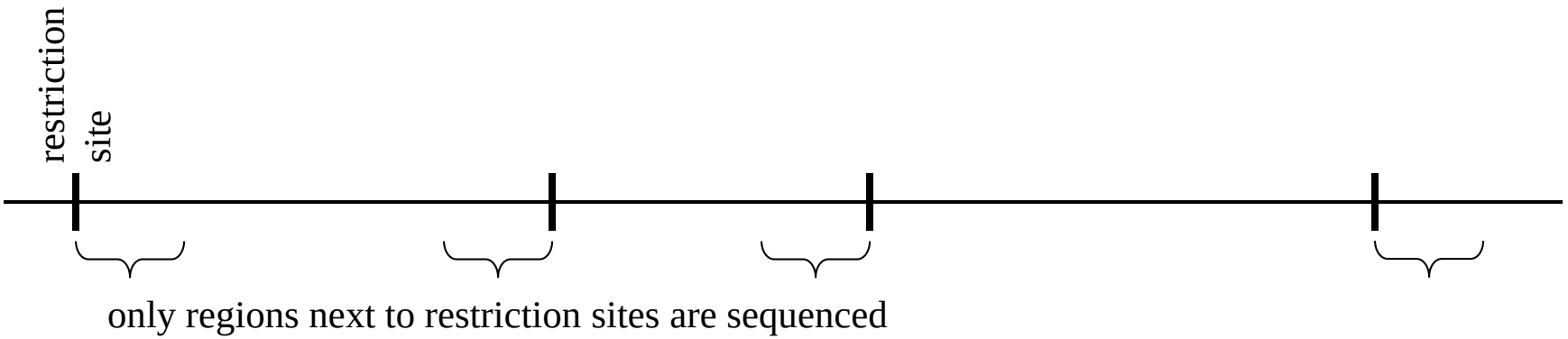
Genotyping by sequencing (GBS)

- 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)



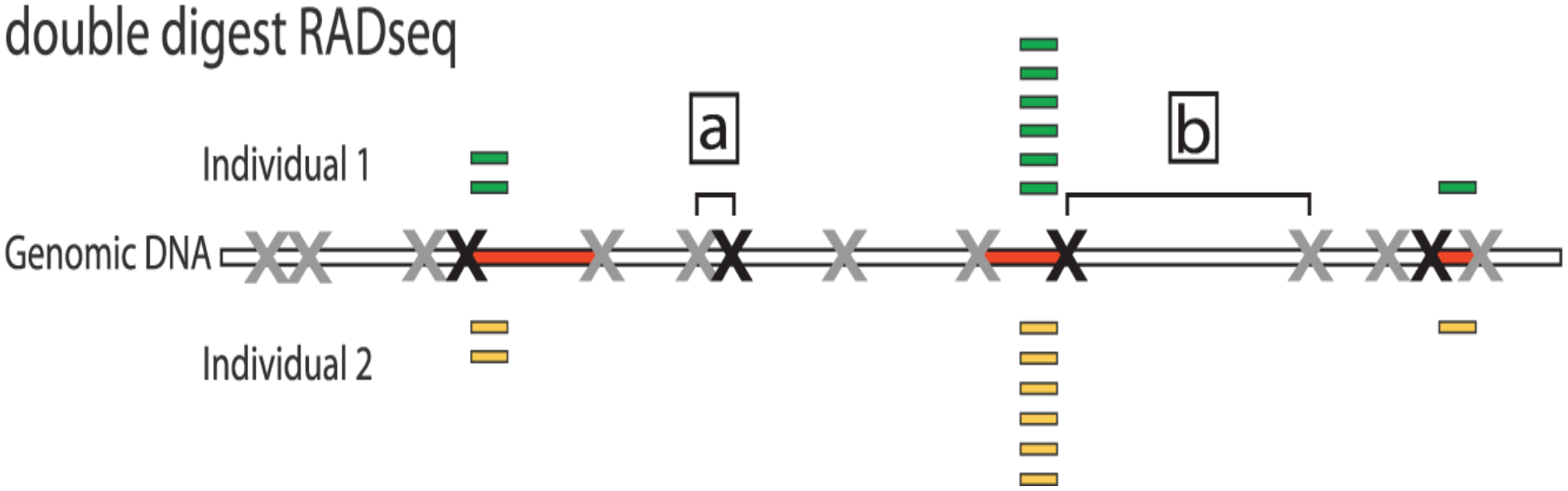
Baird et al, Plos ONE (2009) 3(10): e3376

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)

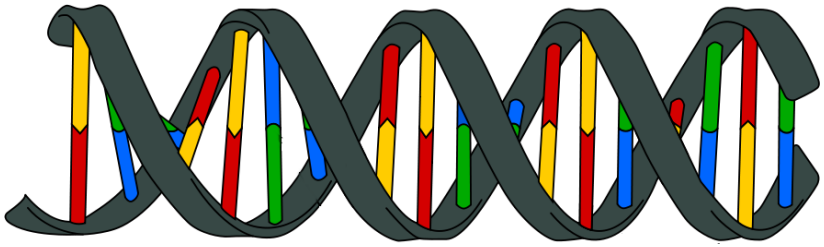
Transcriptomes: RNA-Seq

- 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.

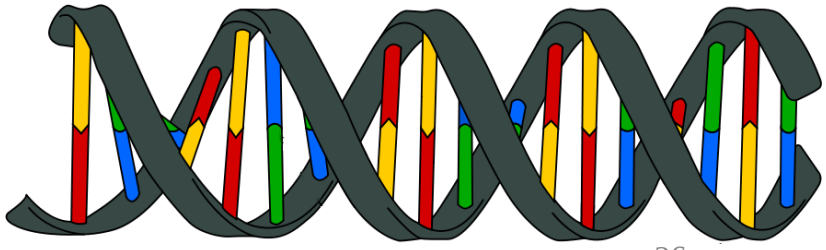
Gene expression

- Measured through transcript (mRNA) abundance

Condition A

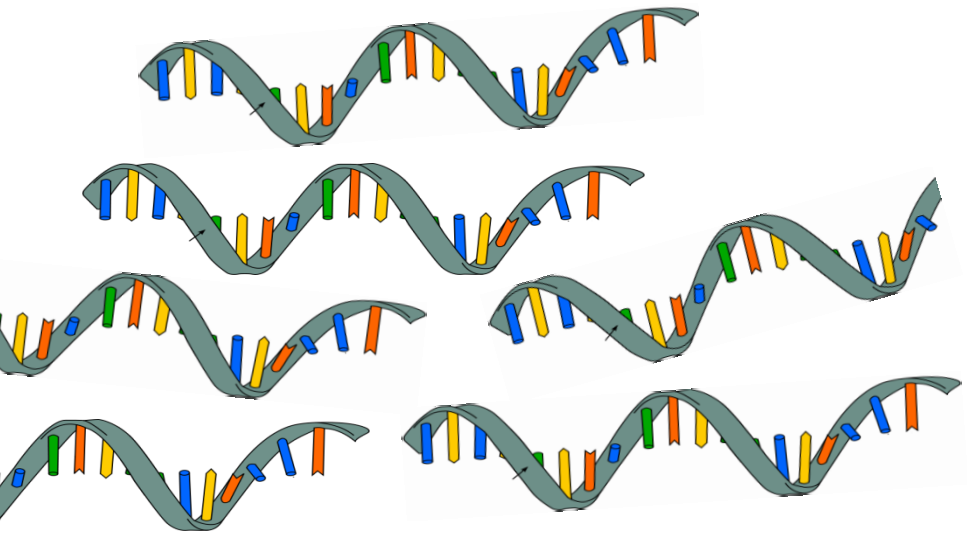


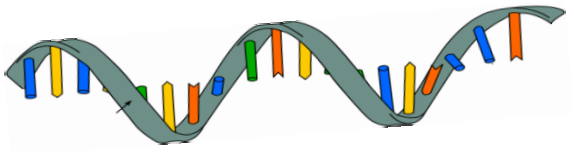
Condition B

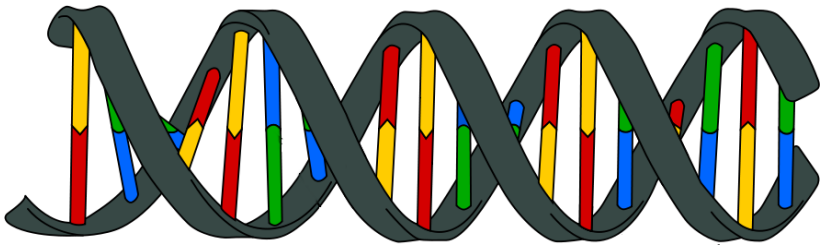


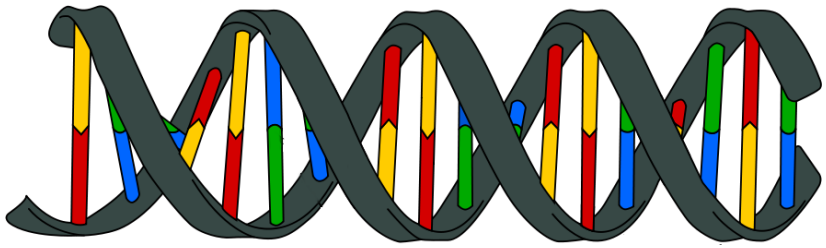
Gene expression

- Measured through transcript (mRNA) abundance





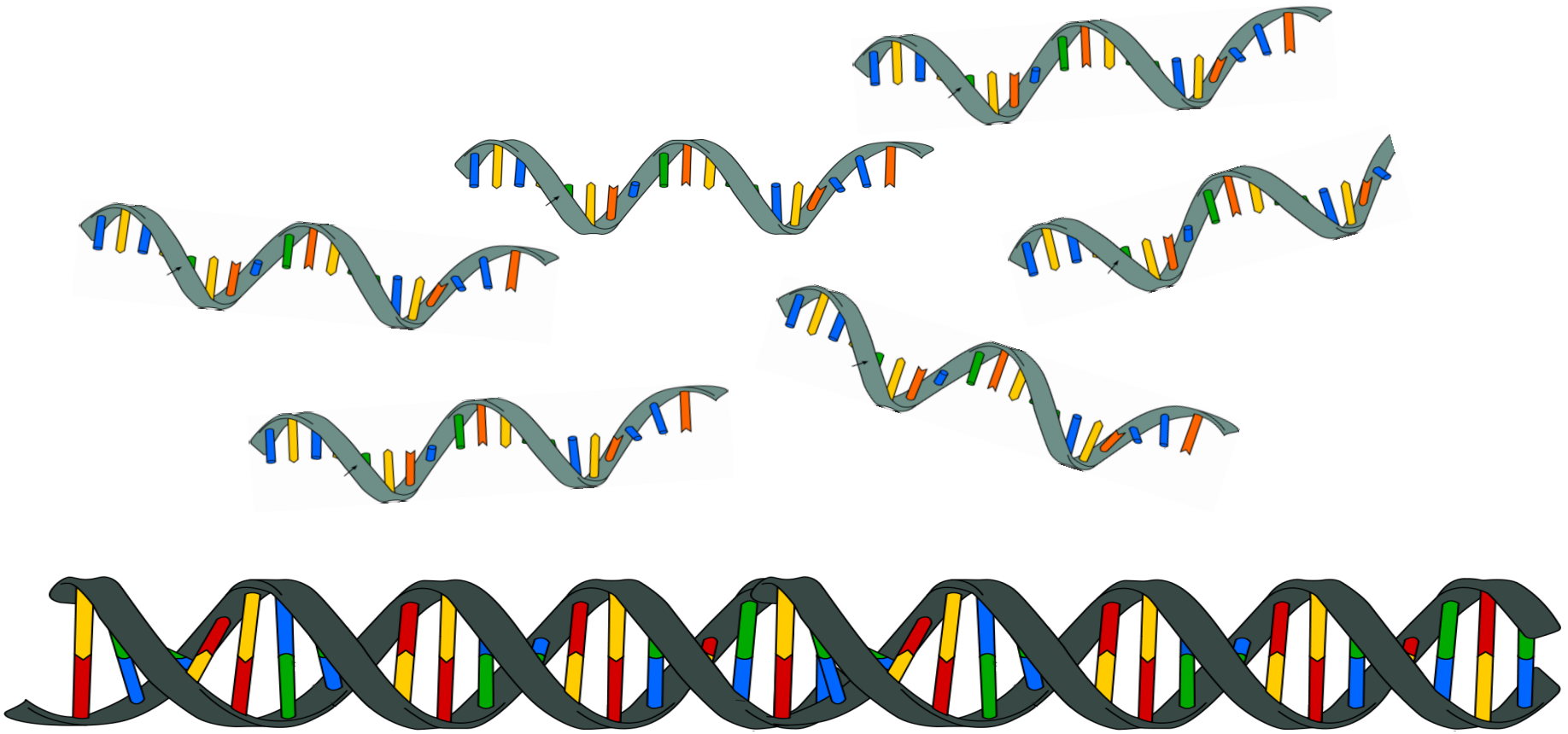




Gene expression: RNA-Seq

- 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

CGGGAAACCAAATCAAGAAGCAGGCAATCCTTAGGATTGCTTTTCGTGGGTAGAGCGAGGGGTTT

ATTTTTCAGTCTTCTCTCGTGGCATTTATTGTCGGTTGGTTTTCTATATATTGCTCGTGCAACTC

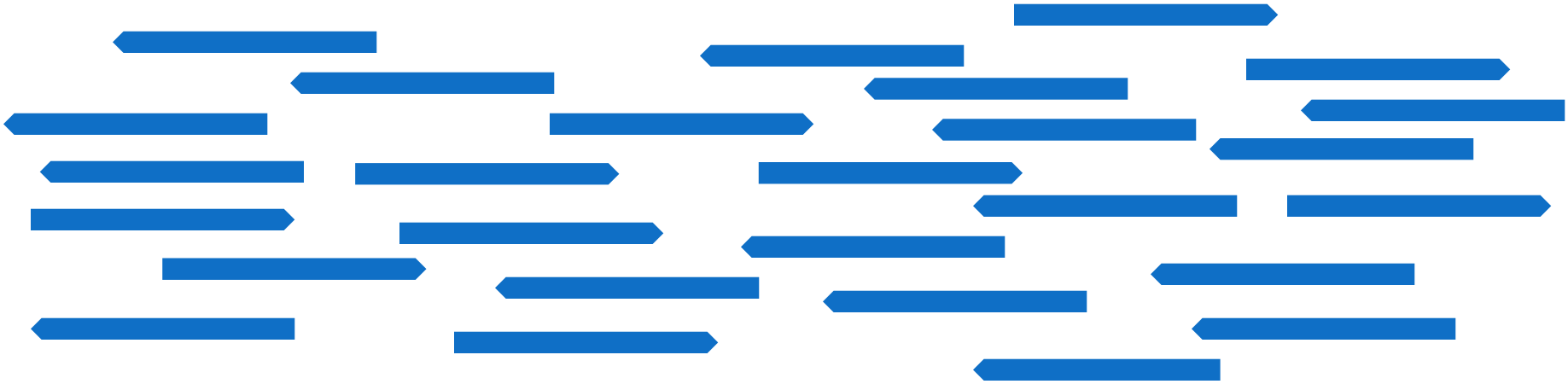
CGTCCCTACCATATCTCATCATCATTATCAATAATATAAGAAACATAATTATCATAATAGAGGAA

CTCTTGCCGGCATTGTGGGCAAAGAGAGAATTGTTGTGTCCACTTCTTGCTCACTTCTTCACACT

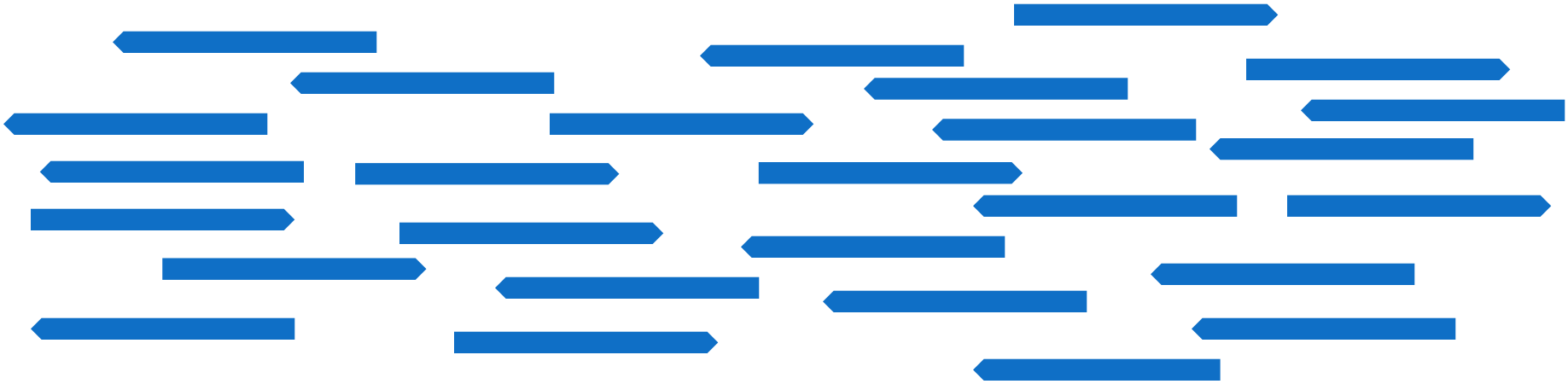
TGTCATAATAACACTCTCTGCTGGTAGAGGTGCAGAATGCTGTAACATACCATCCCCTTCTTTTA

AAAATATATTTCTGGGGATCAATTGACAAAGGATGATATCAAAGTGTACGGATATGTTTCTGAGA

Millions of short sequence reads



Millions of short sequence reads



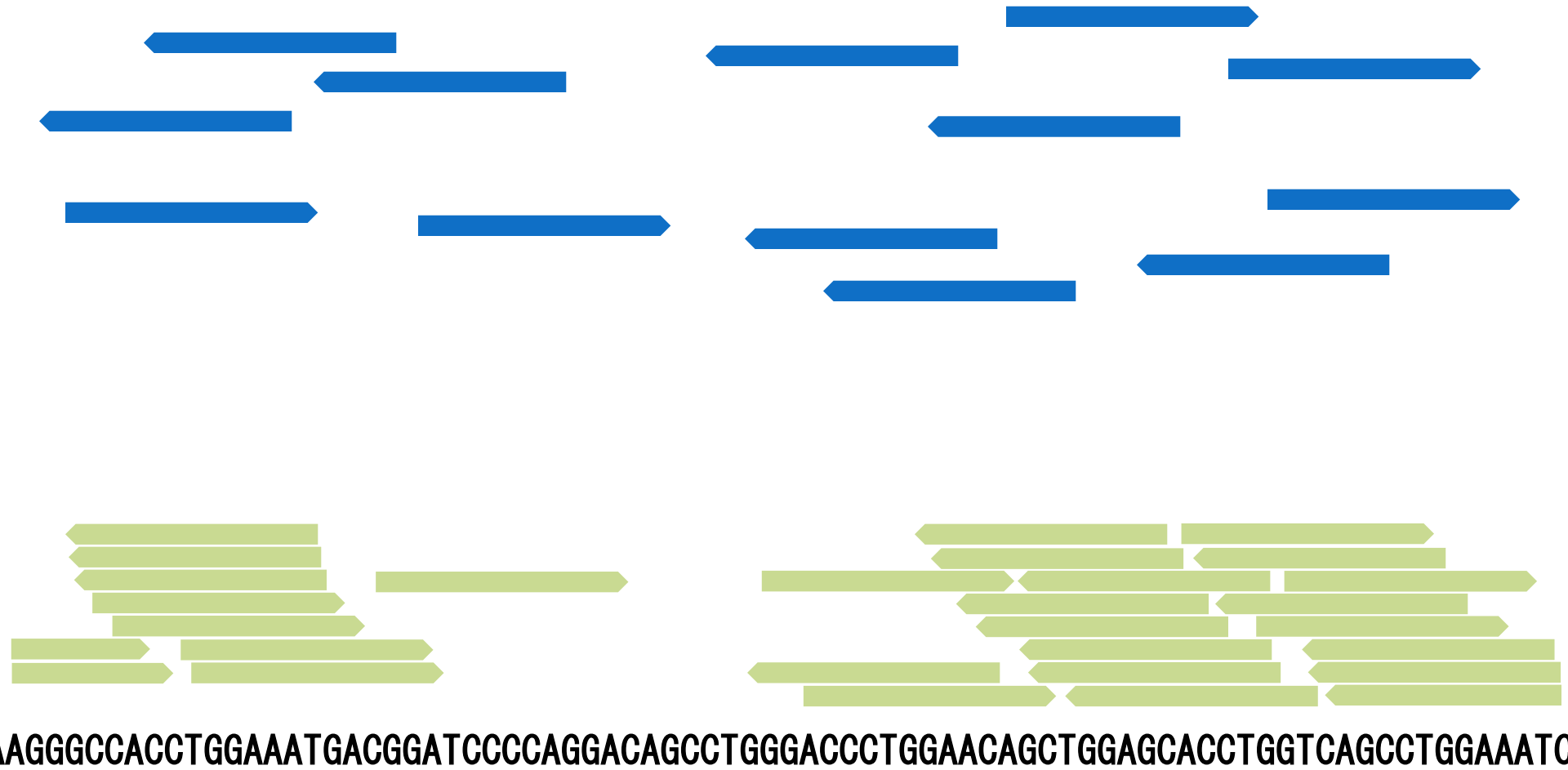
AAGGGCCACCTGGAAATGACGGATCCCCAGGACAGCCTGGGACCCTGGAACAGCTGGAGCACCTGGTCAGCCTGGAAATC

genome sequence

Align short reads to genome

genome sequence

Reads that don’t align in first pass …



AAGGGCCACCTGGAAATGACGGATCCCCAGGACAGCCTGGGACCCTGGAACAGCTGGAGCACCTGGTCAGCCTGGAAATC

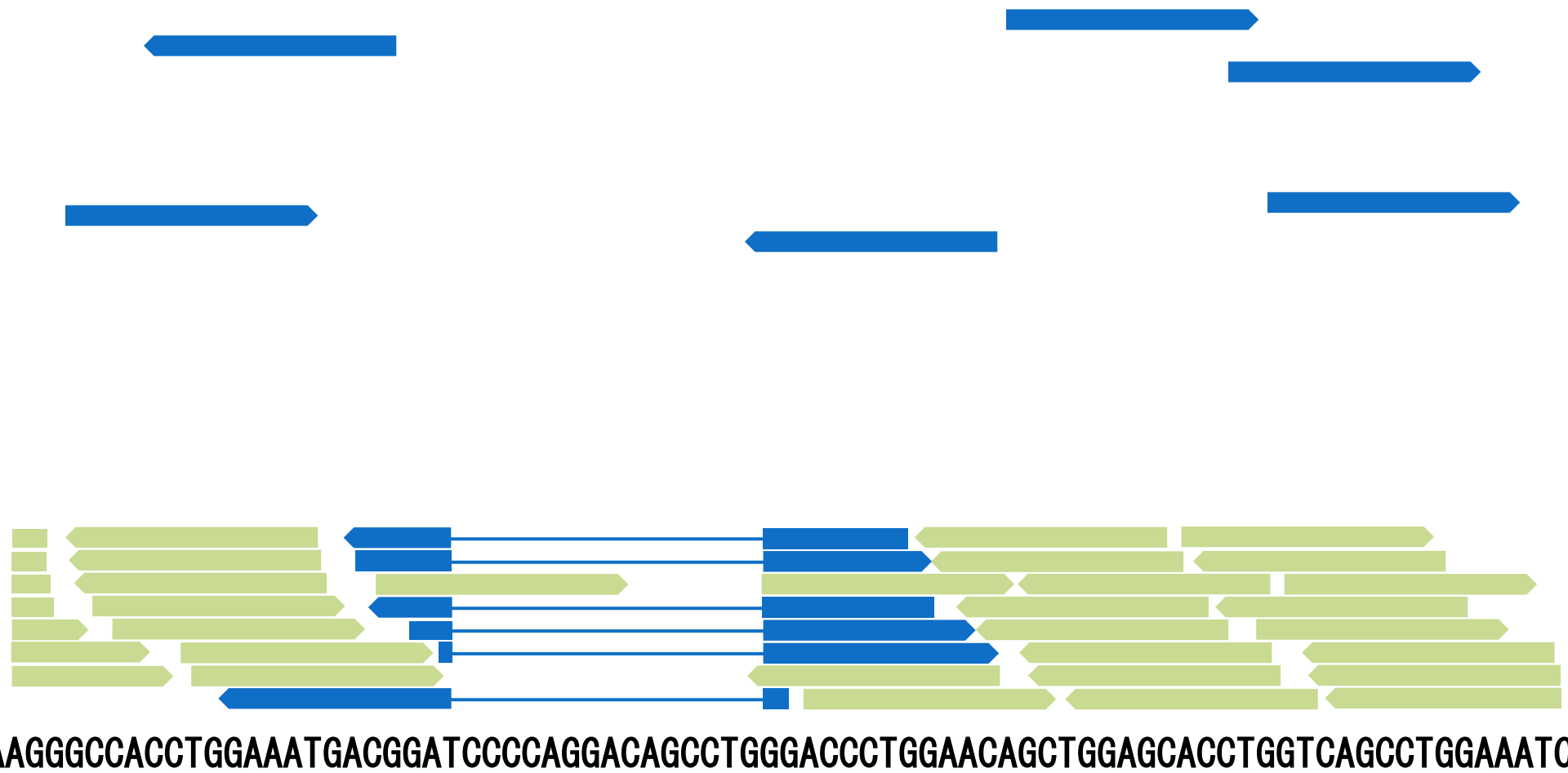
genome sequence

Break into pieces

AGGGCCACCTGGAAATGACGGATCCCCAGGACAGCCTGGGACCCTGGAACAGCTGGAGCACCTGGTCAGCCTGGAAATC

genome sequence

Align allowing for gaps: introns

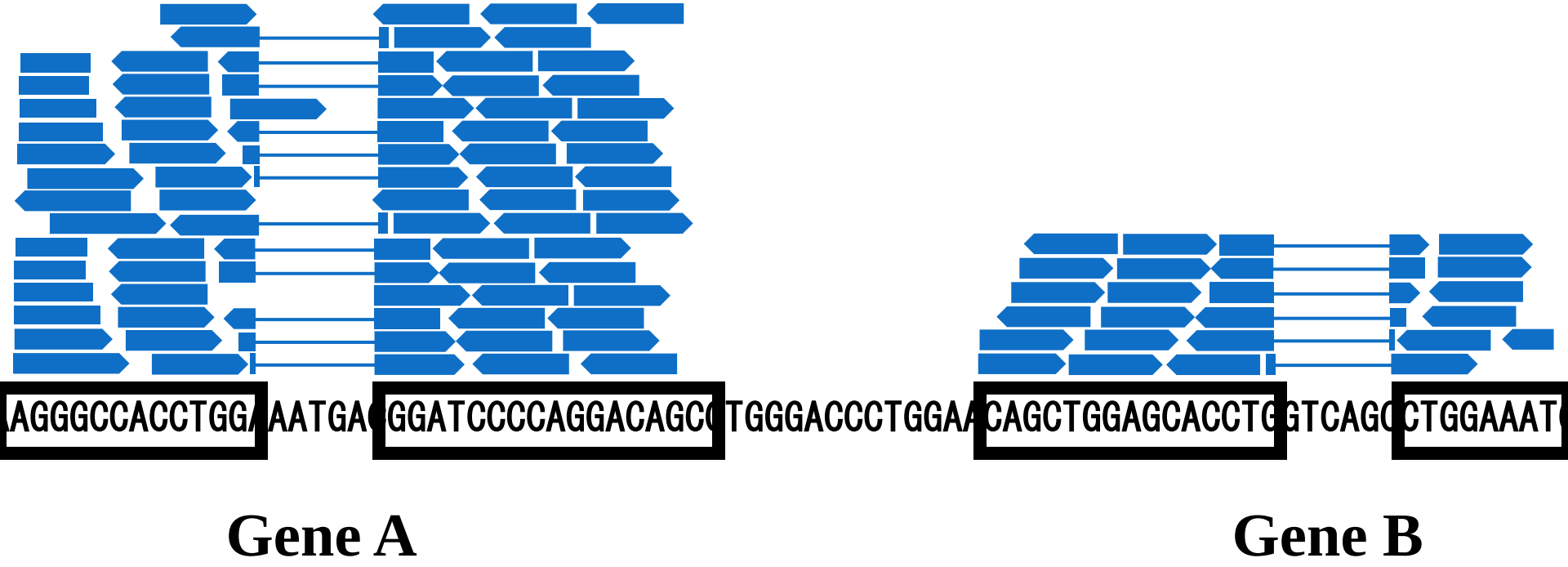


genome sequence

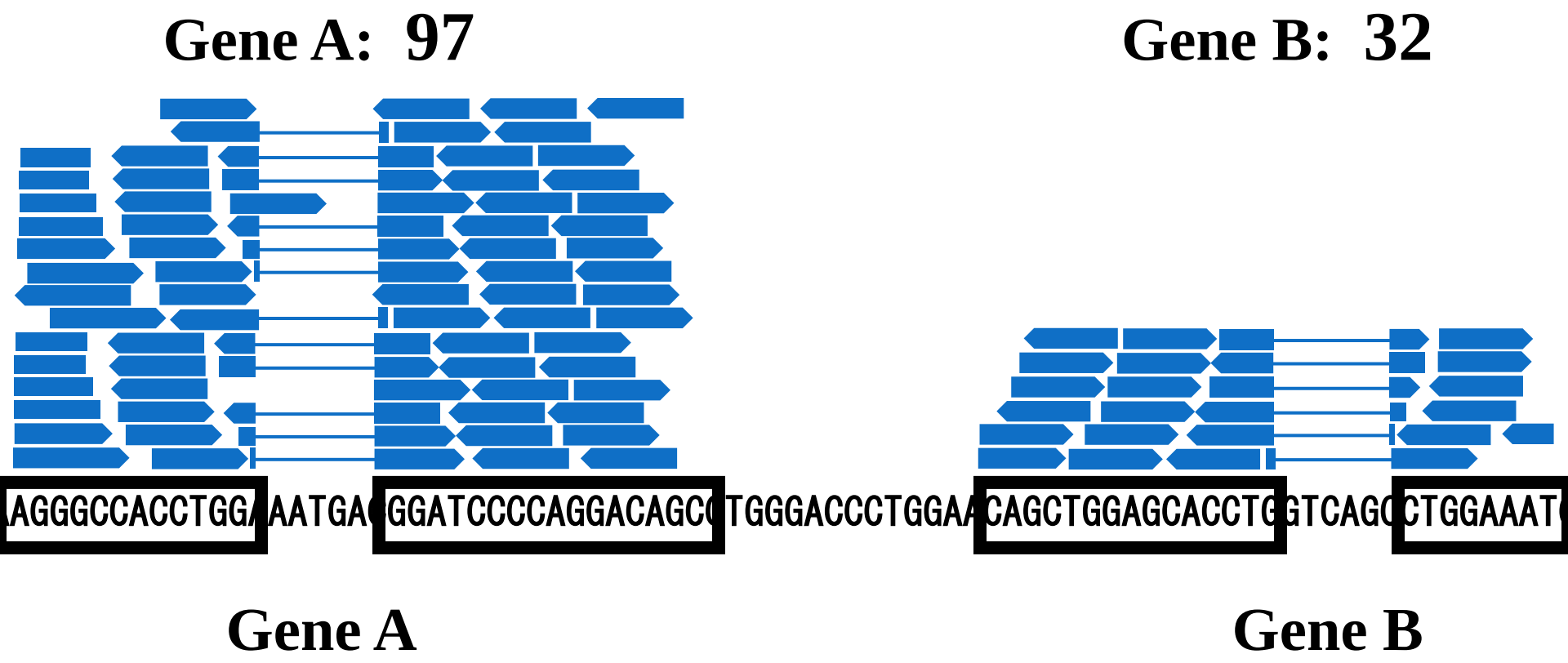
Use alignments to determine which genes contributed which sequenced transcripts

genome sequence

Use alignments to determine which genes contributed which sequenced transcripts



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



Reference sequences

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

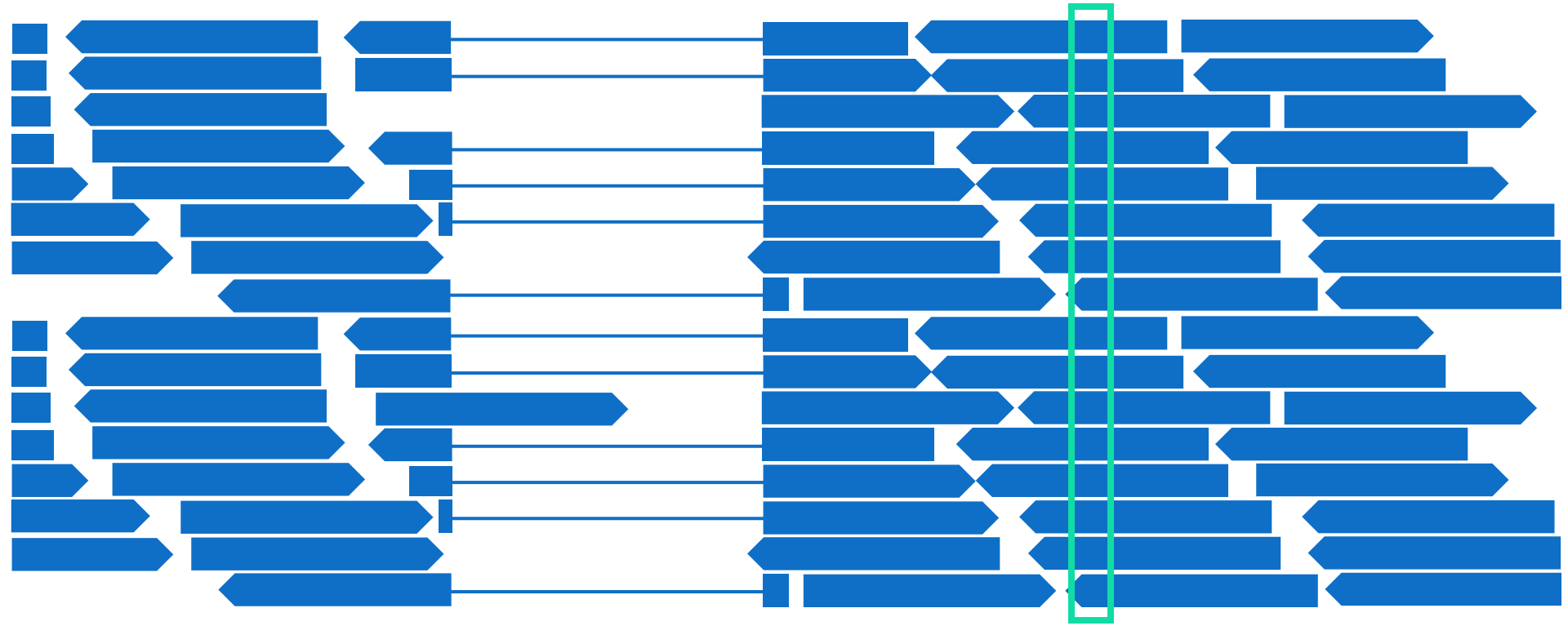
Transcriptome assembly

- 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)

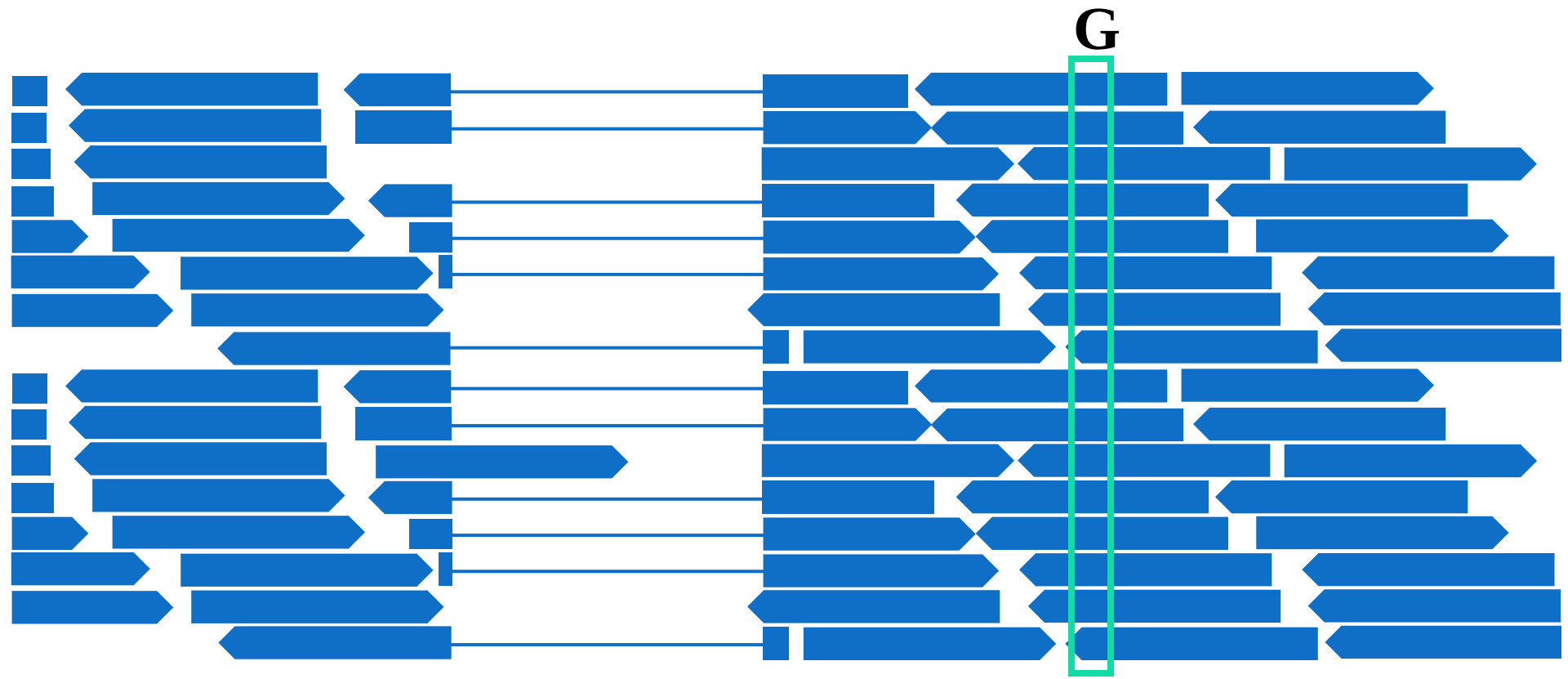
RNA-Seq data

also provides the ability to locate sequence variation across individuals

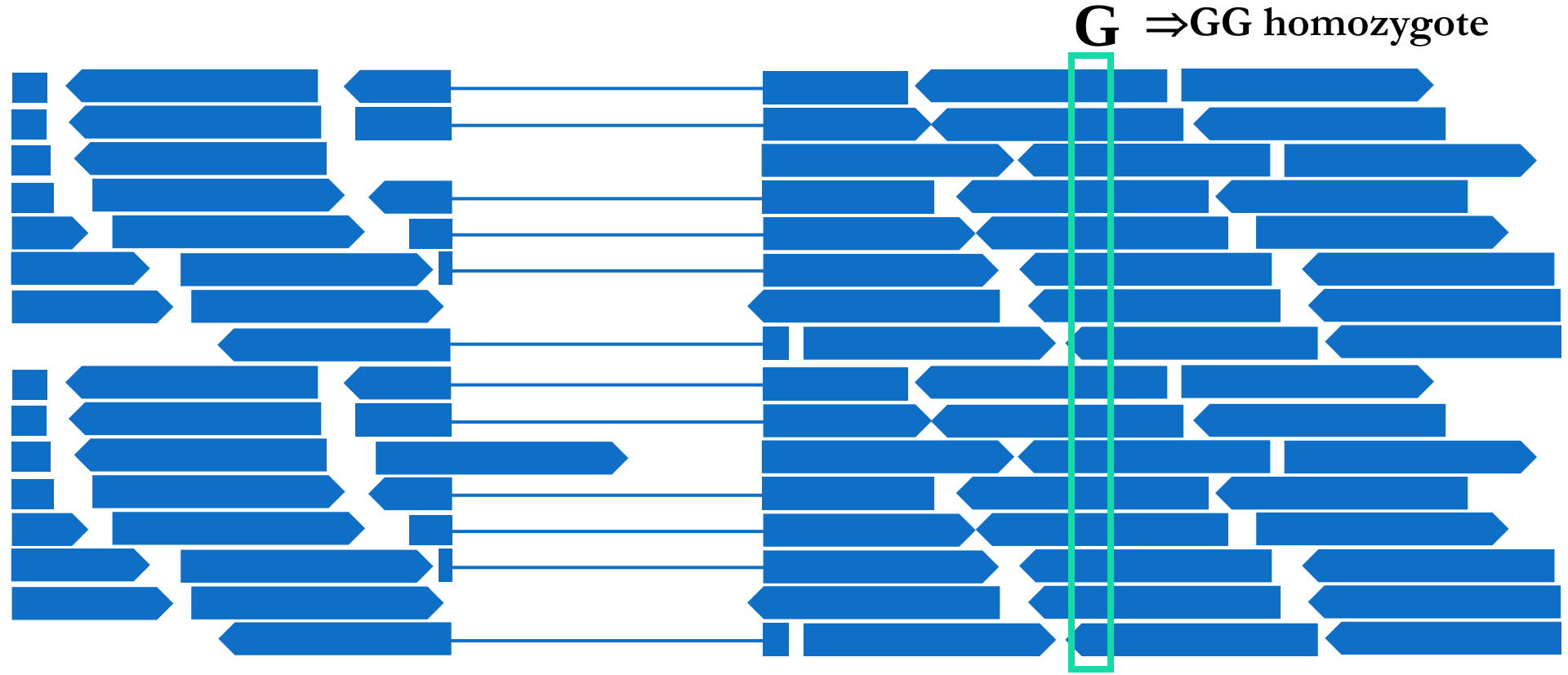
Identifying sequence polymorphisms



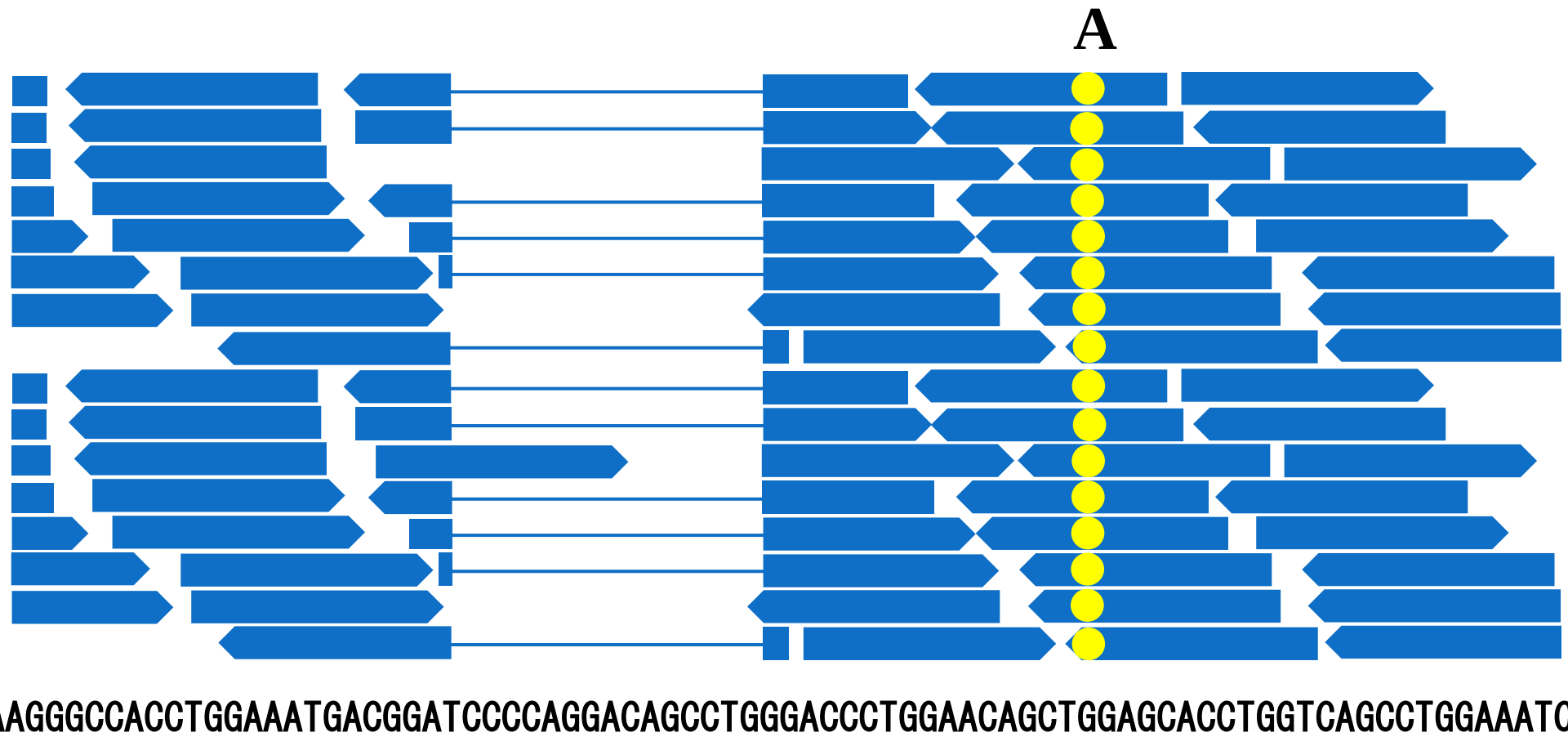
Identifying sequence polymorphisms



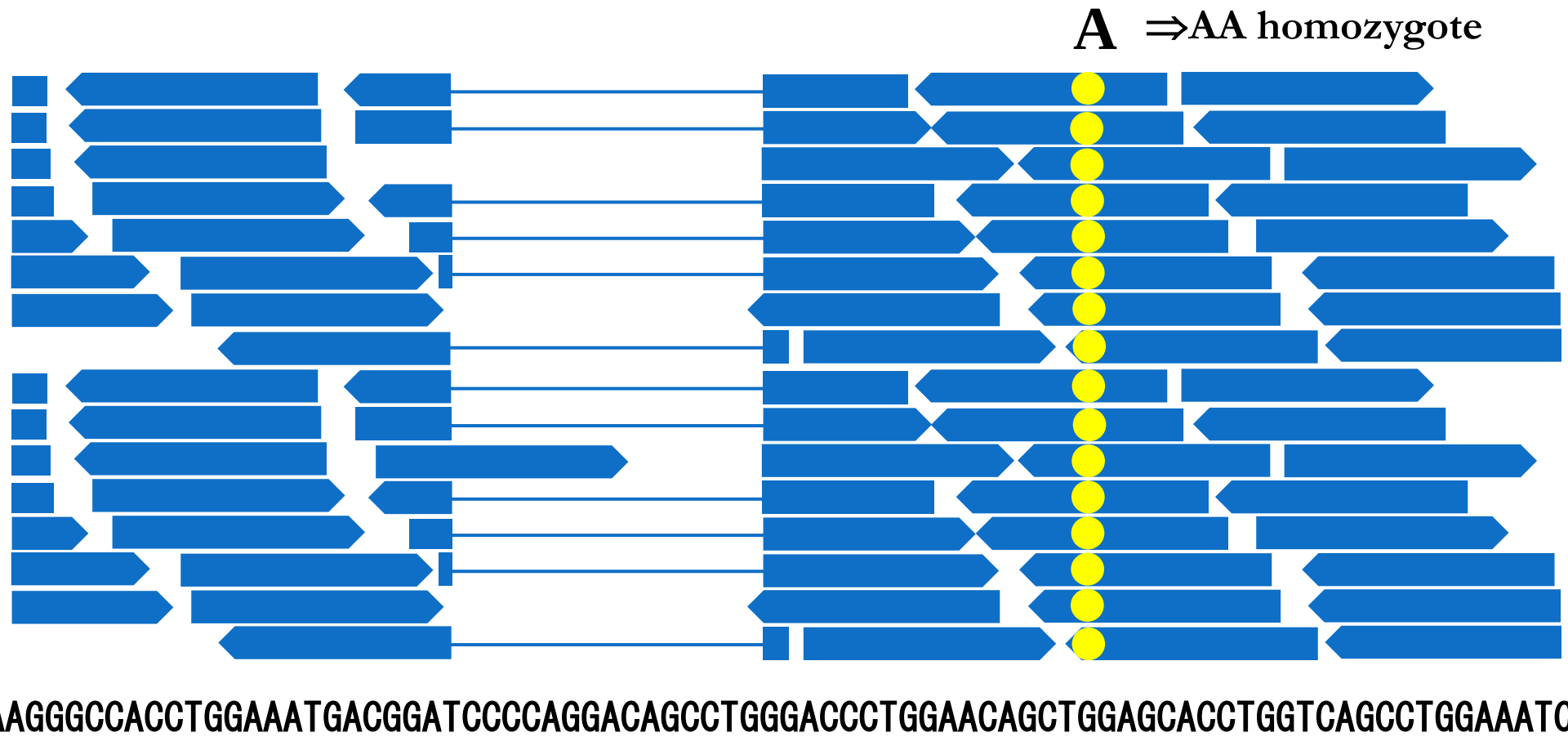
Identifying sequence polymorphisms



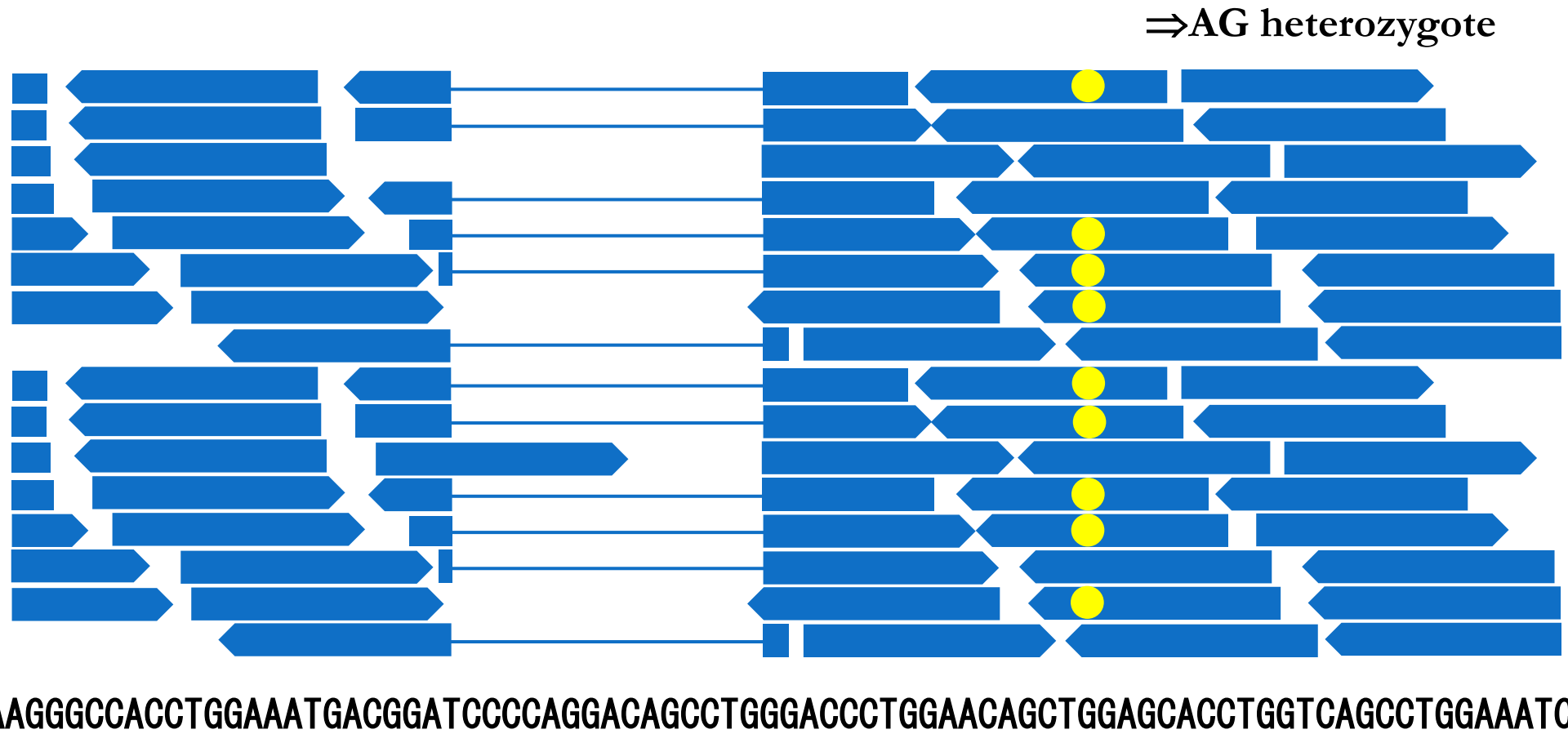
Identifying sequence polymorphisms



Identifying sequence polymorphisms



Heterozygote (~50% of each allele)



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