Introduction to Bioinformatics
and Next Generation Sequencing Data
Part (1)

Tomáš Vinař
Faculty of Mathematics, Physics and Informatics
Comenius University in Bratislava, Slovakia
Outline

- Two-part lecture with brief introduction to NGS and some bioinformatics topics

  - **Lecture part (1):** next generation sequencing, genome assembly and genome projects, comparative genomics (Tomáš Vinař)

  - **Lecture part (2):** sequence alignment and read mapping, more applications of NGS (Broňa Brejová)

- **Workshop for beginners:** NGS file formats (fastq, sam) and read mapping (BB)

- **More advanced workshop:** NGS and comparative genomics (TV, Matej Lexa)
“The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid.”
Translation

Codon (triple of nucleotides) determines 1 amino acid

mRNA: UGGUUGGCUCA
protein: WFGS
# Genetic code

<table>
<thead>
<tr>
<th>Alanine (A)</th>
<th>Isoleucine (I)</th>
<th>Arginine (R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GC*</td>
<td>ATA</td>
<td>CG*</td>
</tr>
<tr>
<td>Cysteine (C)</td>
<td>ATC</td>
<td>AGA</td>
</tr>
<tr>
<td>TGC</td>
<td>ATT</td>
<td>AGG</td>
</tr>
<tr>
<td>TGTT</td>
<td>Lysine (K)</td>
<td>Serine (S)</td>
</tr>
<tr>
<td>Aspartic acid (D)</td>
<td>AAA</td>
<td>TC*</td>
</tr>
<tr>
<td>GAC</td>
<td>AAG</td>
<td>AGT</td>
</tr>
<tr>
<td>GAT</td>
<td>Leucine (L)</td>
<td>AGC</td>
</tr>
<tr>
<td>Glutamic acid (E)</td>
<td>CT*</td>
<td>Threonine (T)</td>
</tr>
<tr>
<td>GAAA</td>
<td>TTA</td>
<td>AC*</td>
</tr>
<tr>
<td>GAG</td>
<td>TTG</td>
<td>Valine (V)</td>
</tr>
<tr>
<td>Phenylalanine (F)</td>
<td>Methionine (M)</td>
<td>Tryptophan (W)</td>
</tr>
<tr>
<td>TTC</td>
<td>ATG</td>
<td>TGG</td>
</tr>
<tr>
<td>TTT</td>
<td>Asparagine (N)</td>
<td>Tyrosine (Y)</td>
</tr>
<tr>
<td>Glycine (G)</td>
<td>AAC</td>
<td>TAC</td>
</tr>
<tr>
<td>GG*</td>
<td>AAT</td>
<td>TAT</td>
</tr>
<tr>
<td>Histidine (H)</td>
<td>Proline (P)</td>
<td>Stop codon (*)</td>
</tr>
<tr>
<td>CAC</td>
<td>CC*</td>
<td>TAA</td>
</tr>
<tr>
<td>CAT</td>
<td>Glutamine (Q)</td>
<td>TAG</td>
</tr>
<tr>
<td></td>
<td>CAA</td>
<td>TGA</td>
</tr>
<tr>
<td></td>
<td>CAG</td>
<td></td>
</tr>
</tbody>
</table>
DNA sequencing

Sequencing technologies produce short reads from random locations in the DNA sample.
DNA sequencing

Position of individual reads on the target DNA is not known

Solved by computational methods:
– **mapping** if target DNA is known
– **assembly** if it is not known
# Overview of Sequencing Technologies

<table>
<thead>
<tr>
<th>Technology</th>
<th>Read length</th>
<th>Errors</th>
<th>Output per day</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1st generation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sanger</td>
<td>up to 900bp</td>
<td>&lt; 2%</td>
<td>3 MB</td>
</tr>
<tr>
<td><strong>2nd (next) generation (cca 2004)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>454</td>
<td>400bp</td>
<td>&lt; 2%</td>
<td>400 MB</td>
</tr>
<tr>
<td>Illumina MiSeq</td>
<td>150bp</td>
<td>&lt; 1%</td>
<td>2 GB</td>
</tr>
<tr>
<td>Illumina HiSeq</td>
<td>150bp</td>
<td>&lt; 0.5%</td>
<td>85 GB</td>
</tr>
<tr>
<td>Ion Torrent</td>
<td>200bp</td>
<td>&lt; 2%</td>
<td>10 GB</td>
</tr>
<tr>
<td><strong>3rd generation (now emerging)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PacBio</td>
<td>up to 14kbp</td>
<td>15%</td>
<td></td>
</tr>
<tr>
<td>Oxford Nanopore</td>
<td>up to 100kbp</td>
<td>30%</td>
<td></td>
</tr>
</tbody>
</table>
**Genome Sequencing Overview**

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1976</td>
<td>MS2 (RNA virus) 40 kB</td>
</tr>
<tr>
<td>1988</td>
<td>Human genome sequencing project (15 years)</td>
</tr>
<tr>
<td>1995</td>
<td>bacterium H. influenzae 2 MB, shotgun (TIGR)</td>
</tr>
<tr>
<td>1996</td>
<td>S. cerevisiae 10 MB, BAC-by-BAC (Belgium, UK)</td>
</tr>
<tr>
<td>1998</td>
<td>C. elegans 100 MB, BAC-by-BAC (Wellcome Trust)</td>
</tr>
<tr>
<td>1998</td>
<td>Celera: human genome in three years!</td>
</tr>
<tr>
<td>2000</td>
<td>D. melanogaster 180 MB, shotgun (Celera, Berkeley)</td>
</tr>
<tr>
<td>2001</td>
<td>2x human genome 3 GB (NIH, Celera)</td>
</tr>
<tr>
<td>after 2001</td>
<td>mouse, rat, chicken, chimpanzee, dog, ...</td>
</tr>
<tr>
<td>2007</td>
<td>Genomes of Watson and Venter (454)</td>
</tr>
</tbody>
</table>
Revolution 1: Shotgun sequencing

Shotgun sequencing:

• Sequence the whole genome

• Trust sequence assembler (provided we have high-enough coverage)

BAC-by-BAC:

• Create BACs (approx. 100 kb)

• Genome mapping / select BACs that cover the whole genome with small overlaps

• Sequence BACs one by one
Why is coverage important?

Coverage of individual bases at $10 \times$ genome coverage with reads of length 1,000.
Why is coverage important?

Expected number of contigs when genome of size 1,000,000 is covered by increasing number of segments of length 1,000
<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1976</td>
<td>MS2 (RNA virus) 40 kB</td>
</tr>
<tr>
<td>1988</td>
<td>Human genome sequencing project (15 years)</td>
</tr>
<tr>
<td>1995</td>
<td>bacterium H. influenzae 2 MB, shotgun (TIGR)</td>
</tr>
<tr>
<td>1996</td>
<td>S. cerevisiae 10 MB, BAC-by-BAC (Belgium, UK)</td>
</tr>
<tr>
<td>1998</td>
<td>C. elegans 100 MB, BAC-by-BAC (Wellcome Trust)</td>
</tr>
<tr>
<td>1998</td>
<td>Celera: human genome in three years!</td>
</tr>
<tr>
<td>2000</td>
<td>D. melanogaster 180 MB, shotgun (Celera, Berkeley)</td>
</tr>
<tr>
<td>2001</td>
<td>2x human genome 3 GB (NIH, Celera)</td>
</tr>
<tr>
<td>after 2001</td>
<td>mouse, rat, chicken, chimpanzee, dog,…</td>
</tr>
<tr>
<td>2007</td>
<td>Genomes of Watson and Venter (454)</td>
</tr>
</tbody>
</table>
Revolution 2: Next-generation sequencing

Cost per raw megabase of DNA sequence

Until 2007: Sanger sequencing
Starting in 2008: next-generation (454, Illumina, SOLiD)
Typical Results of NGS Assembly

- Many **short contigs** that can be further combined to **longer scaffolds** by using **paired reads**

- Some portions cannot be resolved due to **long repetitive sequences**

**Example:** Human chromosome 14, 88 Mbp, 70× coverage
(source: GAGE)

<table>
<thead>
<tr>
<th>Method</th>
<th>Contigs</th>
<th>Errors</th>
<th>N50 corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Velvet (basic de Bruijn)</td>
<td>&gt;45000</td>
<td>4910</td>
<td>2.1 kbp</td>
</tr>
<tr>
<td>Velvet (with scaffolding)</td>
<td>3565</td>
<td>9156</td>
<td>27 kbp</td>
</tr>
<tr>
<td>AllPaths-LG</td>
<td>225</td>
<td>45</td>
<td>4.7 Mbp</td>
</tr>
</tbody>
</table>

N50: reads with this length or longer contain 50% of the genome
here N50 after error correction is shown
NGS pioneers

(first uses of NGS for mapping rather than sequencing novel genomes)

Wheeler, D. A. et al. (Nature 2008) Baylor College of Medicine
The complete genome of an individual by massively parallel DNA sequencing
– genome of James Watson
– sequenced by 454
NGS pioneers
(de novo assembly of a small genome is possible from hybrid data)

Reinhardt J.A. et al. (Genome Research 2009)
De novo assembly using low-coverage short read sequence data from the rice pathogen Pseudomonas syringae pv. oryzae
– genome size 6Mb, N50 scaffold size 500kb
– Illumina 26x coverage, some paired and unpaired 454 reads
NGS pioneers

(first Illumina-only vertebrate-size genome)

Li R. et al. (Nature 2010)

The sequence and de novo assembly of the giant panda genome
– genome size 2.3Gb, N50 contig size 40kb
– paired end Illumina 56x coverage
– different paired libraries up to 1kb, read length 52bp
For comparison: Sanger sequencing

Lindblad-Toh K. et al. (Nature 2005)
Genome sequence, comparative analysis and haplotype structure of the domestic dog
– genome size 2.4Gb, N50 contig size 180kb
– Sanger 7.5x coverage
– paired reads from libraries 4kb-200kb
A typical modern genome paper

- One or more genomes finished to high quality (chromosomes)
- Additional genomes sequenced to draft quality (scaffolds)
- or: 10s of individuals from various population groups to map population diversity
- Basic genome description / annotation
- Advanced analyses (comparative genomics, population genetics, . . . )
- Biological function (individual genes or gene families responsible for specific traits)
Comparative genomics

Nothing in biology makes sense except in the light of evolution
(Theodosius Dobzhansky, 1973)
Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures

Alexander Stark1,2*, Michael F. Lin1,2, Pouya Kheradpour1,2, Jakob S. Pedersen1,2, Leopold Parts1,2, Joseph W. Cartier1, Madeline A. Crosby1, Matthew D. Rasmussen1, Sushmita Roy2, Ameya N. Deoras1, J. Graham Ruby3,4, Julius Brennecke1, Harvard FlyBase curators1, Berkeley Drosophila Genome Project1.

Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10,000 Vertebrate Species

Genome 10K Community of Scientists®

The 1KP Project
The 1000 plants (oneKP or 1KP) initiative is a public-private partnership generating large scale gene sequence information for 1000 different species of plants. Major
Why do we need so many genomes?

- Common features of genomes: Which genes are responsible for basic biological functions?
- Differences between genomes: Which mutations are responsible for typical traits of individual species?
- Identify elusive functional regions. (RNA genes, regulatory regions, ...)
- Study evolutionary mechanisms and their impact on genomes.
Human-mouse comparison
Whole-genome alignments

For each section of reference genome (e.g. human) find corresponding sections of other genomes.

- **Human** AGTGGCTGCCAGGCTG---GGATGCTGAGGCCTTGTTTGCAGGGAGGT
- **Rhesus** AGTGGCTGCCAGGCTG---GGTTGCTGAGGCCTTGTTTGCCGGGAGGT
- **Mouse** GGTGGCTGCCCGGGGCTG---GGTTGCTGAGGCCTTGTTTGCCGGGAGGT
- **Dog** AGTGGCTGCCCGGGGCTG---GGTTGCTGAGGCCTTGTTTGCCGGGAGGT
- **Horse** GATGGCTGCCCAGGGGCTG---GGTTGCTGAGGCCTTGTTTGCCGGGAGGT
- **Armadillo** AGTGGCTGCCCAGGGGCTG---GGTTGCTGAGGCCTTGTTTGCCGGGAGGT
- **Chicken** AGTGGCTGCCCAGGGGCTG---GGTTGCTGAGGCCTTGTTTGCCGGGAGGT
- **X. tropicalis** AATGGCTTCCATTTGTGCGCCGCTGAGGTCTTGTCTTCTGCGGAGAT
Mutation rates vary depending on function

- Neutrally evolving sequence

- **Purification selection** (typical for functional element)
  ⇒ slower than neutral mutation rates (higher similarity)

- **Positive selection** (typical in regions responsible for novel functions)
  ⇒ faster than neutral mutation rates (greater divergence)
Human Accelerated Regions

We are looking for regions, which:

- evolved slowly for a long time (purification selection)
- in Humans evolved surprisingly fast (positive selection)
Human Accelerated Regions: HAR1

Region of length 118

18 differences between human and chimp
2 differences between chimp and chicken

300 mil. years

6 mil. years

Human CGAATTGTGCGCTAGACGACACGTCAAGCGGCCGGAGAATGTTCTAT
Chimp CGAATTGTAGACACATGTCAAGCACAGTGAGAAATAGTTCTAT
Gorilla CGAATTGTAGACACATGTCAAGCACAGTGAGAAATAGTTCTAT
Rhesus CGAATTGTAGACACATGTCAAGCACAGTGAGAAATAGTTCTAT
Mouse CGAATTGTAGACACATGTCAAGCACAGTGAGAAATAGTTCTAT
Cow CGAATTGTAGACACATGTCAAGCACAGTGAGAAACCGTTCTAT
Dog CGAATTGTAGACACATGTCAAGCACAGTGAGAAACAGTTCTAT
Chicken CGAATTGTAGACACATGTCAAGCACAGTGAGAAACAGTTCTAT

28
Whole-genome studies: positive selection in protein coding genes

Looking at patterns of mutation in protein coding genes:

- **Synonymous:** local “neutral” speed
  e.g. ACA (Thr) → ACT (Thr)

- **Non-synonymous:** possible functional changes
  e.g. ACA (Thr) → AAA (Lys)

High ratio of non-synonymous to synonymous changes (ω) is a sign of **positive selection**
IGF1R: Example of a gene under positive selection

Marmoset Genome Consortium, Nature Genetics, 2014
Summary

• Next-generation sequencing is a powerful tool for many applications

• Nowadays, it is “easy” to sequence genomes (draft)

• More difficult to obtain high quality genomes (assembled to the level of chromosomes) and annotate / analyze data

• Comparative genomics benefits from large availability of genomes (even draft quality is often useable if we have a good reference)

• Whole-genome comparative studies can sometimes yield unexpected results
Introduction to Bioinformatics
and Next Generation Sequencing Data
Part (2)

Broňa Brejová
Faculty of Mathematics, Physics and Informatics
Comenius University in Bratislava, Slovakia
Recall from part (1)

- NGS technologies produce large number of reads from a given DNA sample
- Some technologies produce short reads with low error rates; others long reads with high error rates
- Reads can be used to assemble genomes
- In comparative genomics we compare groups of related genomes

Part (2)

- Sequence alignment and read mapping
- More applications of NGS
Sequence alignment, homology search

- Given two sequences/sequence databases, find regions of similarity between them
Input: two sequences

ggccttggagttgactgtcctgctccttt
gagggcatttctcagagagaggaagtggccctcaca
ttttaatccgcctttccacagccttgtctcttttc
cagacccatgggagaggggctgaggggtg
tggctgagccccaccaagtacgcgtactct
gcaggtccctctcccccaagggccgtggcccttg
ggagcccggtggatccagtagtgagctacgcctcc
accccccgccctactcgggcagtttaacccctt
gtgtcttcacttgacagacatcgtgaaaccgagcc
cggcccgacgagaagggccataaatgacctatgt
gtccagctttctaccatgccttttcagggagcgc
agaggttacccgagccggccagggcaggccctc
cgcgcgccccaccgccgaattgcccgccgtgcctc
cctcgcctcccggtcctcactcctattttcttttgcc
agacggcagttgcctctctccaatggaagcc
acccccagctccctt...

tgatgcccagggatgttgccgttcagacatccgg
cagagaaagtctcatccacctacgtgggcacctacctac
tatcactacttttagcaaaactcaagcagggac
ngtgcagggcataagcgtatcggttaagggtgg
cggcattgcctggagagacgacaaatggttcc
acgactacgagaactttccaaagcgtctgctc
aagttcatgagaacagaccatccagtccgctggg
cgagccggagttgcaaaaactctgcgtgccggcggc
tccaagggcagttggccctaattctcaactaac
ccgaccatcgaagaagccgcccaagtttgtgga
aaagggcaacctcggaggtgccctcttttccaccc
tgcatcctcagactgcggccacaacaaccagaag
ccctacacaccccaaagggcaagatgatttttc
ggacatcacaagggcttgagcgcgctcggaga
aggcccgacgcaagcagcgaatttgccctgccc
gaggagctcatccg...
Output: similar regions in the form of alignments

ggccctttgagttgactgtacctgctgctccttt
gagggccatttccagagagagagagaatgggcctca
ttttatccgccttccccacagcctttgctctttc
cagaccccatggagagggaggggtcgaggggtg
tggctgagccccacccaggtacgcgtcactct
gcaggtccctctcccccaaggccgtgaggttg
gagggcctggatccagctgagtgagtcgcctcc
ccccctgctactcgggcagtttaaccctgtgttt
tcacttacctaccatgctccttttcagagacgc
agaaggtaccgagcagggccaggccctctcc
tgcggacgaggaagtggttggtctctctctctgg
tcagcggcagtgggcctctctcactaagtggat
ctgatgcggaggtgtttcggtcacgcatccgg
acgagaaagttcatcacctactctggttggtggt
tgcgccatggagacagcagaaagttggtttctggtc
gagactacgagaaaacttcgctggccggcg
tcaaggacagtggccagtttctctcaactac
cgcacccatcgagaagccgccaaaagttttgtga
aaagggcaacctcagaggtgctctctctctctccccc
pgcaggtaaccagcgagcagggcagggcagcggcctctca
tcgccgcagagggcaattgccgccgctgctctc
tccgtgtccctgctctcactctctttctcttgct
tagacggcaggtggccctctctcacaactggagcc
accccccagctcccct...

tcccgacgacgaaggccataatgacctatgtg
tccagctttctaccatgctccttttcagagacgc
agaaggtaccgagcagggccaggccctctcc
tgcggacgaggaagtggttggtctctctctctgg
tcagcggcagtgggcctctctcactaagtggat
ctgatgcggaggtgtttcggtcacgcatccgg
acgagaaagttcatcacctactctggttggtggt
tgcgccatggagacagcagaaagttggtttctggtc
gagactacgagaaaacttcgctggccggcg
tcaaggacagtggccagtttctctcaactac
cgcacccatcgagaagccgccaaaagttttgtga
aaagggcaacctcagaggtgctctctctctctccccc
pgcaggtaaccagcgagcagggcagggcagcggcctctca
tcgccgcagagggcaattgccgccgctgctctc
tccgtgtccctgctctgctctcactctctttctcttgct
tagacggcaggtggccctctctcacaactggagcc
accccccagctcccct...

CCCGACGAGAAGGCCATAATGACCTATGTGTCAGCCTTCTACCATGCGCTTT
|| || || || || || || || || || || || || || || || || || ||
CCGGACGAGAAGTCCAT---CACCTACGTGGTCACCTACTATCATCCTACTTT
Sequence alignment, homology search

- Given two sequences/sequence databases, find regions of similarity between them
- Display in the form of an alignment

```
CCCGACGAGAAGGCCATAATGACCTATGTGTCCAGCTTCTACCATGCCCTTT
||||| ||||| ||||| ||||| |
CCGGACGAGAAGTCCAT---CACCTACGTGGTCACCTACTATCACTACTTT
```

Insert dashes (gaps) so that corresponding bases in the same column. A good alignment has many aligned matching bases, few gaps.
What are alignments good for?

- **Read mapping:**
  From which part of the genome is a given read? Must be fast, mapping many reads.

- **Determine function (e.g. of a protein):**
  Similar sequences often have the same or similar function.

- **Comparative genomics/evolution:**
  Search for homologs, sequences which have evolved from the same common ancestor.
  Ideally, gaps correspond to insertions and deletions, aligned bases to conserved bases and substitutions.
Many types of alignment: local/global/…

What task we want to solve?

- **Global alignment**: Align two sequences across their whole length
  e.g. two homologous proteins

- **Local alignment**: Find similarities between shorter regions of long sequences
  E.g.: find homologous genomes between two related genomes

- **Read mapping**: Align entire read to a short region of the reference genome

- **Whole-genome alignment**: Select representative subset from all local alignments between two genomes
Many types of alignment: pairwise vs multiple

- **Pairwise alignment**: create alignments with two rows

  CCCGACGAGAAGGCCATAATGACCTATGTGTCCAGCTTCTTCTACCATGCCTTT
  || || || || || || || || || || || || || || || || || || || || ||
  CCGGACGAGAAGTCCAT---CACCTACGTGGTCACCTACTATCACTACTTT

- **Multiple alignment**: align more than 2 sequences

  Usually global or whole-genome alignments

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>ctccatagcaatgt-cagagataaggcagagcggat--------ggtggtgac</td>
</tr>
<tr>
<td>Rhesus</td>
<td>ctccatggcaatgt-cagagatagggcagacgcggat--------gctggtgac</td>
</tr>
<tr>
<td>Mouse</td>
<td>ttt--tgacaaca--tagagac-tgagatagaaat---------atgctgac</td>
</tr>
<tr>
<td>Dog</td>
<td>-tccccgctaatgtacaagatggggcag-gaaga--a-----tgtgctgaa</td>
</tr>
<tr>
<td>Horse</td>
<td>-tccacggcaatac-tggagatggggcagagcaga--agat-ggtgatgaa</td>
</tr>
<tr>
<td>Armadillo</td>
<td>ctgcataagaaatct-cagagatgggggaagcaga--------agacattcat</td>
</tr>
<tr>
<td>Opossum</td>
<td>atccatggaacat-cagaggtgggagaatagaaga----tggaatgaa-</td>
</tr>
<tr>
<td>Platypus</td>
<td>acccgggggggg-aagaggaagggccggccg---------------</td>
</tr>
</tbody>
</table>
Speed of alignment algorithms

- **Exact algorithms** find the best alignment possible.

- **Pairwise:** dynamic programming algorithms.
  - Running time grows with the product of the sequence lengths.
  - Practical only for moderately long sequences (e.g. proteins, RNAs)

- **Multiple:**
  - Running time grows exponentially with the number of sequences.
  - Practical only for few very short sequences.
Speed of alignment algorithms

- **Exact algorithms** find the best alignment possible, but are often too slow in practice.

- **Heuristic algorithms** may miss some alignments, but have more practical running time.

- **Main trick: alignment seeds**

  BLAST starts by finding all exact matches of length 11 between input sequences; these are called hits. Hits can be found fast (e.g. by hashing, BWT, ...). Each hit is extended to a full alignment by slower methods.
Speed vs. sensitivity

- **Alignment seed**: e.g. $w = 11$ consecutive matches

  **Alignments without a seed are not found**

  **Sensitivity** of a tool: what portion of real alignments are found

  $\text{CCCGACGAGAAGGCCATAATGACCTATGTGTCAGCTTCTACCATGCCTTT}$

  $\text{CCGGACGAGAAGTCCAT---CACCTACGTGGTCACCTACTATCACTACTTT}$

  - **Increasing $w$**: fewer hits, faster program, less sensitive
    - **Decreasing $w$**: more hits, slower program, more sensitive

  - More complex alignment seeding used in many tools
Example: sensitivity in random alignments

Use \( w = 11 \), random alignments with probability \( p \) of match
(human-mouse: \( p \approx 0.7 \))

Sensitivity: probability that alignment contains 11 consecutive matches
Examples of popular alignment tools

- Water: exact pairwise local alignment
- Needle: exact pairwise global alignment
- FASTA, BLAST, BLAT, Lastz, Last, . . . : heuristic pairwise local alignment
- Clustal, muscle, mafft, t-coffee, . . . : heuristic multiple global alignment
- TBA/multiz, Pecan/Enredo, . . . : multiple whole-genome alignment
- Bowtie, BWA, SHRiMP, SOAP, Last, . . . : heuristic read mapping
- New tools published all the time!
More on read mapping

- Different tools/settings depending on technology:
  for short low-error reads (e.g. Illumina) can use longer seeds
  for longer high-error reads (e.g. PacBio) needs shorter seeds, more
  similar to genome alignment

- Unified format for storing alignments: sam/bam

- Specialized tools for some applications (e.g. split mapping for
  structural variants and RNA-seq)

- What if a read has multiple matches?
Mappability

- Repeats and duplicated sequences cannot be distinguished with short reads (paired reads can help)
- Mapping quality: reflects probability that alignment is wrong (read comes from a different part of the genome)
- Report all high-quality alignments or choose one randomly (e.g. to estimate coverage in duplicated areas)
Example: Details of Bowtie mapper

B. Langmead et al. (Genome Biology 2009)
Ultrafast and memory-efficient alignment of short DNA sequences to the human genome

- builds index of reference in memory
- then aligns reads one after another
- optimized for high speed, low memory, but fewer features
- aligns 25 million 35-bp reads per hour
- 2.2 GB of memory for the human genome (2.9 GB for paired-end)

Indexing based on Burrows-Wheeler transform (BWT, 1994) and Ferragina-Manzini index (FM index, 2000)
Genome resequencing, population genomics

Explore differences between individuals by sequencing multiple genomes from the same species

- Discover single nucleotide polymorphisms (SNPs)
  - map reads to reference, identify differences

- Discover structural variants
  - bigger differences, but harder to find
Discovering structural variants

Inconsistent read pairs, split alignments, coverage variation

From Eichler, Nature Reviews 2011
**Genome resequencing, population genomics**

Explore differences between individuals by sequencing multiple genomes from the same species

**Association studies:**

- Correlate SNPs with diseases/other traits
- Find causal variants

**Population history:**

- Study ancient population sizes, migrations, domestication etc.
- Purely from present day individuals or ancient DNA (up to hundreds thousands year old)
Example: effective population size of human populations

Li and Durbin (Nature 2011) Inference of human population history from individual whole-genome sequences

Compare the two copies of a chromosome from the same individual
Locate blocks with different density of heterozygous sites
Estimate time to most recent ancestor; longer times $\Rightarrow$ larger populations
Cancer genomes

Find mutations occurring in cancer cells
– causes of disease
– therapeutic targets
– disease variants (large variability)

Typical sample genetically heterogeneous
– clonal structure explored by single-cell sequencing

Large projects

The Cancer Genome Atlas (TCGA)
Cancer Cell Line Encyclopedia (CCLE)
Metagenomics/environmental genomics

Sequence DNA of uncultured microbial community
– may have more than 10,000 species
Also sequence DNA fragments from environment (e.g. sediments, ice)

Explore diversity by read binning
assign reads to taxonomic groups by similarity to known genomes

From: Willerslev et al. (Nature 2014) Fifty thousand years of Arctic vegetation and megafaunal diet.
Human microbiome

Human body: $10^{13}$ human cells, $10^{14}$ bacterial cells

The Human Microbiome Project Consortium (Nature 2012)
Structure, function and diversity of the healthy human microbiome

242 people, 15-18 sites per person, some repeat visits
Large variation within and between individuals
RNA-seq

- Sequence RNA extracted from a sample, map reads to genome
- Measure expression level of individual genes
- Discover new genes, splicing variants, polymorphisms, RNA editing, transcript fusion in cancer
Epigenomics

- Expression of genes is regulated by DNA methylation and histone modifications
- ChIP-seq with antibodies specific for individual modifications
- DNase-seq: DNase I Hypersensitivity Site footprinting – discovers open chromatin structure (potential regulatory sequences)
Chromatin immunoprecipitation (ChIP)

Cross-link whole cells with formaldehyde

Isolate genomic DNA

Sonicate DNA to produce sheared, soluble chromatin

Add protein-specific antibody

Immunoprecipitate and purify immunocomplexes

Reverse cross-links, purify DNA and prepare for sequencing

From E.R. Mardis (Nature Methods 2007)
Molecule interactions

- ChIP-seq: DNA-protein interactions (transcription factors)
- HITS-seq: RNA-protein interactions
- Hi-C, ChIA-PET: DNA-DNA interactions in the nucleus

From: Lieberman-Aiden et al. (Science 2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome
ENCODE project (Encyclopedia Of DNA Elements)

Human genome, also some model organisms (mouse, drosophila, worm)
Many NGS technologies, e.g. RNA-seq, ChIP-seq in many tissues
Segmentation of genome to regions based on experimental data

UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs, Comparative Genomics)

H1-hESC Genome Segmentation by Combined Segway+ChromHMM from ENCODE/Analysis

TSS Bright Red Predicted promoter region including TSS
PF Light Red Predicted promoter flanking region
E Orange Predicted enhancer
WE Yellow Predicted weak enhancer or open chromatin cis regulatory element
CTCF Blue CTCF enriched element
T Dark Green Predicted transcribed region
R Gray Predicted Repressed or Low Activity region
Summary

- Next-generation sequencing is a powerful tool for many applications, including genome (re)sequencing, transcriptomics, epigenomics, interactions

- NGS data requires sophisticated bioinformatics analysis

- ZOO of tools for read mapping and alignment

Next: workshops

- **Workshop for beginners:** NGS file formats (fastq, sam) and read mapping (BB)

- **More advanced workshop:** NGS and comparative genomics (TV, Matej Lexa)
Common formats: FASTA

Used for storing DNA/RNA/protein sequences

Only name + sequence, name starts with >

NACTACTGTAGAAGTCAAGATTATTCCACAGGATCATATATTCATATGGAGATCAATCTGGTTC

ACTGCCAGCAGAAAAATCTCCCAATTTTTTCCACCAGATATTCGCTGTTGAGT

>sp|P00410|COX2_YEAST Cytochrome c oxidase subunit 2
MLDLLRLQLTTFIMNDVPTPYACYFQDSATPNQEGILELHDMFYLILVGLSWMLYTIVMT
YSKNIAYKYIKHGQTIEVIWTIFAPAVILLIAIFPSFILLYLCDEVISPAMTIKAIGYQWLYKW
EYSDFINDSGETVEESYPVIPDELEEGQLRLLDTDTSVMVVPDTHIRFVVTADVHDFAI
SGIKVDATPGRNLQVSAIQLQREGVFYGACSELCGETGTHANMPIKIEAVSLPKFLEWLNEQ

>sp|P21534|COX2_SCHPO Cytochrome c oxidase subunit 2
MLFFNSLNDAPSSWALYFQDGASPSVLGTVHNDYLMTFIFIGVIYAIKAVIEYNNSH
PIAAKYTHGSIIVEFIWLTIPALILILVALPSFKLLYLDEVQKPSMTVKAIGRQFWYWYN
FVTFNENEPVSFSYMPEEDLEEGSLRQLEVDNRLVLPIDTRIRLTSGSVHVWSVPSLGIK
CDCIPGRNLQVSLISIDREGLFYGCSELCGVLHSSMPIVVQGVSLIREDLAWLEEN
Common formats: FASTQ

Used for storing reads including quality values for each base

@: read name, technology-specific format

@HWI-ST1218:80:D0VGUACXX:4:1101:1321:1960 1:N:0:CAGATC
NACTACTGTAGAAGTTCAAGATTATTCCACAGGATCATCATATTCCATATGGAGATCAATCTGGTTC
+
#1=D?DDDFHHHIGHIIIIIIIIIIIIIIIGIIIIIIGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII

@HWI-ST1218:80:D0VGUACXX:4:1101:1321:1960 2:N:0:CAGATC
ACTGCCAGCAGAAAAATCTCCCAATTTTTTACCAGATATTTTGGCCGCTTCAGACTGATTTGAAAGT
+
CCCFFFFFFHHHHFIGIIJIIJJJJJIIJJJIIIIJIIIIIIJIIIIIIIGFEIIIIIIJJJJJGJHFHHH

**Base quality** \( q \): probability of error \( 10^{-q/10} \)

Quality encoded as a single character with ASCII code \( q + 33 \)

For example symbol \( + \) has ASCII value 43,

which means \( q = 10 \) and probability of error \( 10^{-1} = 10\% \)
## Base quality codes

<table>
<thead>
<tr>
<th>q error</th>
<th>q error</th>
</tr>
</thead>
<tbody>
<tr>
<td>! 0 1</td>
<td>2 17 0.02</td>
</tr>
<tr>
<td>&quot; 1 0.794</td>
<td>3 18 0.0158</td>
</tr>
<tr>
<td># 2 0.631</td>
<td>4 19 0.0126</td>
</tr>
<tr>
<td>$ 3 0.501</td>
<td>5 20 0.01</td>
</tr>
<tr>
<td>% 4 0.398</td>
<td>6 21 0.00794</td>
</tr>
<tr>
<td>&amp; 5 0.316</td>
<td>7 22 0.00631</td>
</tr>
<tr>
<td>, 6 0.251</td>
<td>8 23 0.00501</td>
</tr>
<tr>
<td>( 7 0.2</td>
<td>9 24 0.00398</td>
</tr>
<tr>
<td>) 8 0.158</td>
<td>: 25 0.00316</td>
</tr>
<tr>
<td>* 9 0.126</td>
<td>; 26 0.00251</td>
</tr>
<tr>
<td>+ 10 0.1</td>
<td>&lt; 27 0.002</td>
</tr>
<tr>
<td>, 11 0.0794</td>
<td>= 28 0.00158</td>
</tr>
<tr>
<td>- 12 0.0631</td>
<td>&gt; 29 0.00126</td>
</tr>
<tr>
<td>. 13 0.0501</td>
<td>? 30 0.001</td>
</tr>
<tr>
<td>/ 14 0.0398</td>
<td>@ 31 0.000794</td>
</tr>
<tr>
<td>0 15 0.0316</td>
<td>A 32 0.000631</td>
</tr>
<tr>
<td>1 16 0.0251</td>
<td>...</td>
</tr>
<tr>
<td>Z 57 2e-06</td>
<td></td>
</tr>
</tbody>
</table>
Common formats: SAM/BAM

Used for storing results of **read mapping** (alignments)

**SAM**: text format, (somewhat) human readable

**BAM**: binary format, less space, convert to SAM via samtools

Header plus one row for each read

**Columns**: read ID, flag, contig, position, mapping quality, CIGAR, 3 columns for paired reads, read sequence, read quality, optional fields

**Flags**: binary encoded, e.g. if forward or reverse strand

**CIGAR**: matches, insertions, deletions, introns, etc., e.g. 87M2D43M

One row of a SAM file:

```
HWI-ST1300:156:H7599ADXX:1:1110:17800:29845 145 contig0004     \
10 50 41M contig0346 339 0     \
CTATAGATCTTATTACCGCTTATCCAAAGCTGAAAGTGACT  \\
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:41 YT:Z:UU NH:i:1
```