Bioinformatik

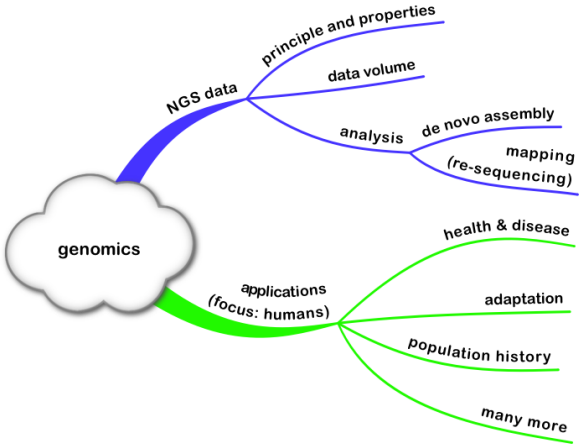
Stefanie Hartmann

Wintersemester 2019 / 2020, Universität Potsdam

(Human) genomics

Dec 06, 2019

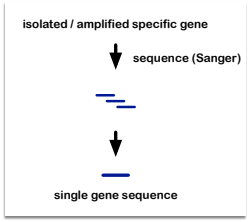
Today's topics



Analysis of selected genes

one globin sequence

- isolate / amplify sequence
- and/or identify in databases
  - by keyword
  - by sequence similarity
- further analyse
  - compare with other globins from the same species
  - compare with globins from different species
  - analyze evolutionary divergence, adaptation, identify conserved functional aspects, etc
  - integrate with other data (e.g., gene expression)



More data!

	in humans
selected (protein-coding) genes	~500 bp (aa) ~1,500 bp (cDNA) ~25,000 bp (gDNA)
mitochondrial genomes	~16,500 bp (37 genes, no introns, very little intergenic regions)
one transcriptome	~21,000 protein-coding genes
one (contemporary) nuclear genome	each: ~3,000,000,000 bp (~98% noncoding, ~21,000 genes on 23 pairs of chromosomes)
many contemporary nuclear genomes	
extinct nuclear genomes	

- generation of nuclear genome data
- applications of genome data
- focus on human genomes

Example: human genome

Sanger / dideoxy sequencing (1975)

- two human genome sequences (2001)

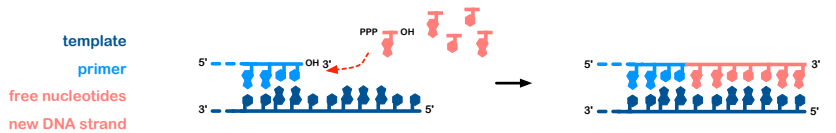
Next generation (since 2005)

- 454, Illumina, Helicos, SOLiD3, Ion Torrent, PacBio, ...
- different read lengths & errors, cheaper, more data per run, higher sensitivity
- thousands of human genomes (& transcriptomes)

data analysis? questions?

- one genome, many genomes

Sequencing (general)



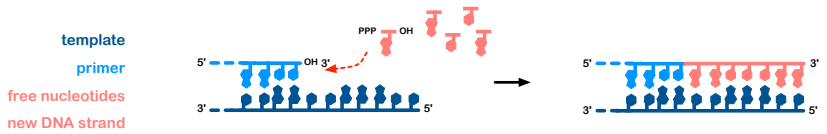
sequencing by synthesis from a template (most platforms)

new nucleotides are incorporated

- chemical signal
- information about corresponding base
- conversion into digital signal

only short stretches can be sequenced at a time

(most) NGS technologies



Sanger

- one template at a time
- reads of 500-800bp
- expensive!
- low-throughput

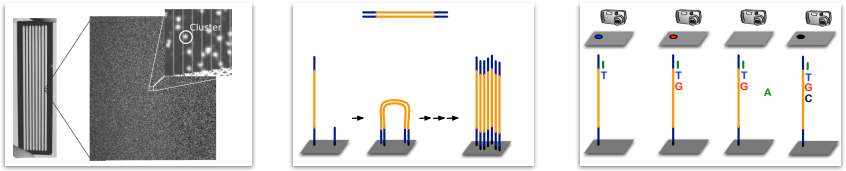
NGS: Illumina

- many templates at a time
- high-throughput
- reads of 75-150bp
- more sensitive
- much cheaper!

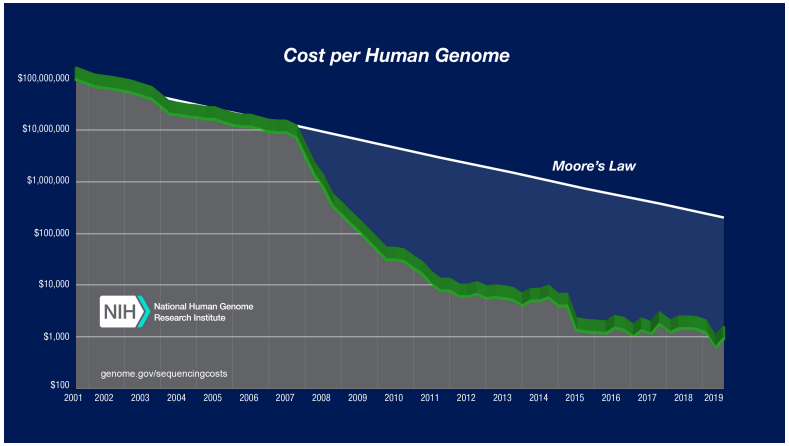
Illumina

	iSeq 100 System	MiniSeq System	MiSeq Series	NextSeq Series
Run Time	9–17.5 hours	4–24 hours	4–55 hours	12–30 hours
Maximum Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb
Maximum Reads Per Run	4 million	25 million	25 million †	400 million
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp

	NextSeq Series	HiSeq 4000 System	HiSeq X Series‡	NovaSeq 6000 System
Run Time	12–30 hours	< 1–3.5 days	< 3 days	~13–25 hours (dual S1 flow cells) ~16–36 hours (dual S2 flow cells) ~44 hours (dual S4 flow cells)
Maximum Output	120 Gb	1500 Gb	1800 Gb	6000 Gb
Maximum Reads Per Run	400 million	5 billion	6 billion	20 billion
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp



https://youtu.be/fCd6B5HRaZ8



https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

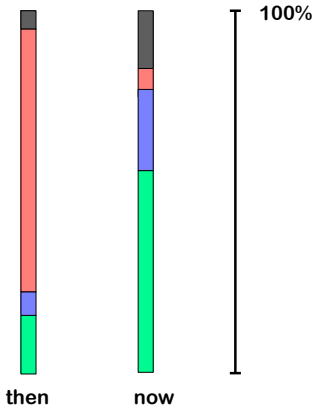
The real cost of a genome

sample collection, experimental design

sequencing

data management & storage

downstream analyses

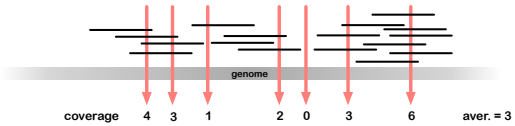


Sequence data

- very short (Sanger: ~500bp, most NGS: 75-150bp)
  - average human gene (coding): 1,500 bp
  - human mitochondrial genome: 16,500 bp
  - human chromosomes: 50,000 bp to 250,000 bp
- required
  - handling huge amounts of data
  - assembly into genes / chromosomes/ genomes
  - genome annotation (gene finding)

de novo assembly

- sample preparation
  - tissue from one organism (multiple cells)
  - isolate, fragment, prepare DNA for sequencing
- input for assembly
  - millions of sequence reads
  - each base in the genome is covered multiple times
  - ➡ coverage (depth): average number of times each base is covered by independent reads



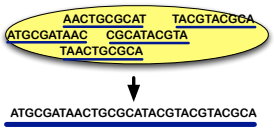
example
genome: 1000bp
read length: 100bp
30 reads ➜ 3x coverage

de novo assembly

assumption
- identical overlapping sequence reads come from the same genomic location

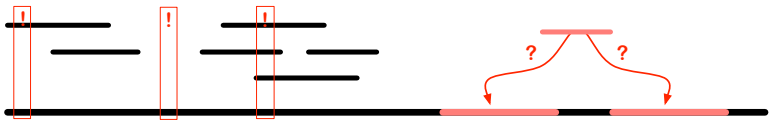
algorithms
- graph-theoretic approaches
- different variations & implementations

output:
- assembled fragments (scaffolds)
- far from perfect



de novo assembly

challenges
- huge data volumes
- sequencing errors
- genome regions not (sufficiently) sequenced
- duplicated genome regions
- ...



de novo assembly

challenges
- huge data volumes

$$\text{number of reads} = \frac{\text{coverage} * \text{genome size}}{\text{read length}}$$

$$30,000,000 = \frac{5 * 3,000,000,000}{500}$$

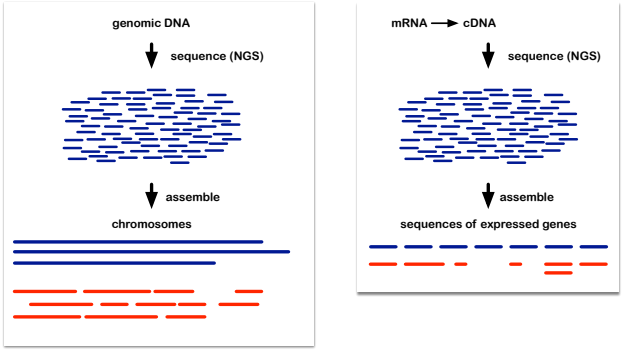
$$1,000,000,000 = \frac{50 * 3,000,000,000}{150}$$

C Venter:

27,271,853 sequence reads of
543 bp average length; generated in
9 months.
5.11 x coverage. The assembly took
20,000 hours of CPU time on
40 machines with
4 processors each.
result: 54,061 sequence scaffolds (6,094)

February 2001

de novo assembly



reconstruction of the target sequence, i.e., the original DNA molecule(s)

- in theory: entire chromosomes or genes
- in practice: unordered contigs, scaffolds

The first (human) genome sequences

CCACCACCACTACCACAATCTCCTCCTCCACCACCACCACCTAGTCACCGGAAACCAATTCATGCCCTAGATTAAGAACCAGAATCCTTTTG
AGACAAGATCTATATATACACTCTACACATTTACTTATACACTTTATCTAGATATAGATGTGTGTGTGTATAAAGGTTCTTTTTCATTACTT
GTTGGTTTAGTGATAAAAAGGAGGAGAGTTTTCAGTTGGGTTTTGGTTTCTTGGTGAGAGAAGAGAGAAAAAGAGAGGATGGGAAGAGGGAA
AGTTGAGCTGAAGAGGATAGAGAACAAGATCAATAGACAAGTTACTTTTGCAAAGAGAAGAAATGGTTTGCTCAAGAAGGCTTATGAGCTTT
CTGTCCTTTGTGATGCTGAGATTGCTCTTCTCATTTTCTCTAACCGTGGCAAGCTCTACGAATTCTGCAGCAGCCCTAGgtgaaaactcttt
aatctttgatatatctacttttttccctcaaagtttctatattttgatccatgttagggttttgattcttaaggattataaaggatatgaaaa
aaagagtgagattaagtgaacgatgatcaaagtacgtacaaagcgtctaagttgctctagatgatgattaaagctagatcaagacccttgaa
ttattttacctaatcagaaatacttaacatttgtttgtgcttctttattgttttgatataattagcagTCATTACAATCACAATCCTGCAAA
TGCAACCAACTCTGCAACAACATCACAGAATGTTAATGGATTCTTCCCTGGATGGATGGTCTGAAAACAATACATATCTATATATGTACACA
TATGTGTGTTTTATAGATCAAAGATCATTGCTTGTGTCATTTGATTTTTGCCAGAGAAACATAGTTCATTATTACAAAACTTTGTTATATAA

- gene finding & annotation
  - computationally find genes
  - computationally predict their function
- gene inventory
  - number? size? structure? location?
- comparison with close and distant species
- ...

The first two human genome sequences

symbolic significance

structure & function of the genome

- only about 25,000 protein-coding genes
- discovery of novel non-coding RNAs
- only 1 - 1.5% of the genome is protein-coding
- about half of the human genome derives from repeats / transposable elements
- many non-coding elements are conserved
- many domains (and domain-architectures) are shared between human and fly/worm/yeast

medicine: disease genes have been identified

new approaches, algorithms, ways to think about & share data

Motivation for more (human) genomes?

basic and applied science

- genome structure & function
- evolution, adaptation, migration, genetic diversity
- health & disease related research

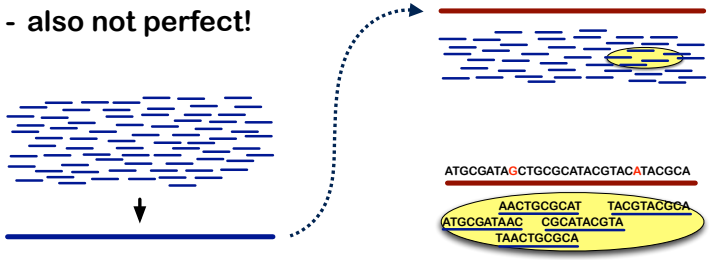
sequenced humans

- from different geographic locations
- healthy, diseased
- genomes, transcriptomes, exomes
- families, unrelated individuals

many non-human genomes

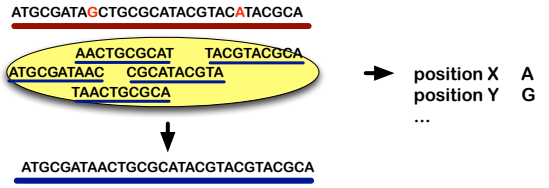
Genome sequencing with a reference

- one high-quality human genome sequence is available
- it can serve as a reference for the new data!
  - computationally less demanding
  - requires lower coverage
  - also not perfect!



Genome re-sequencing

- input
  - sequence reads AND reference genome
- principle
  - align (map) reads to the reference
- output
  - reference-based assembled genes/genome OR
  - positions that differ between reference and data

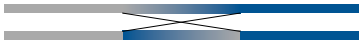


The most interesting parts of the data?

genetic differences between genomes

- **SNPs**
  (single nucleotide polymorphism)
- **indels**
  (insertion or deletion)
- **CNVs**
  (copy number variants)
- **structural variants**

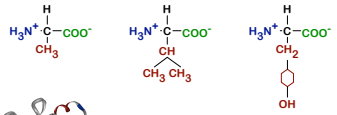
metadata!



Functional changes

adaptive changes, changes relevant for health & disease

- of amino acids with different biochemical properties



- within conserved 3D structures

- within conserved sequences

- within genes that are known to be involved in human diseases

OMIM
Online Mendelian Inheritance in Man
Johns Hopkins University
http://www.ncbi.nlm.nih.gov/omim

Evolutionary analysis

not (just) the functional changes!

- most SNPs are neutral, without functional effects
- abundant & useful markers for evolutionary analyses

questions

- genetic variation within and between populations?
- human origins in each geographic area?
- ancestral migrations: route? individuals? timing?

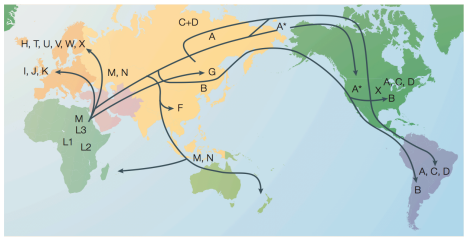
analysis

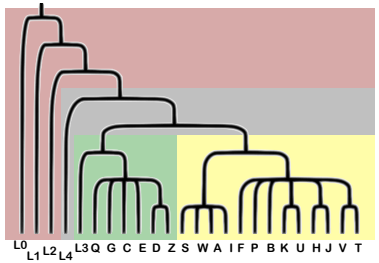
- genetic diversity, phylogenetic analysis, statistical methods (e.g., PCA), dating, correlation with other information, etc

Evolutionary analysis: mtDNA

haplogroups

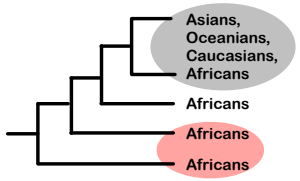
- sex-linked markers (mtDNA, Y-chromosome)
- closely related haplotypes (genetic differences inherited together)
- correlation with geography

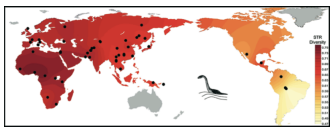


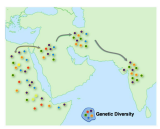


Evolutionary analysis

- the highest genetic diversity is found within Africans
- *H. sapiens* originated in Africa ("out of Africa")

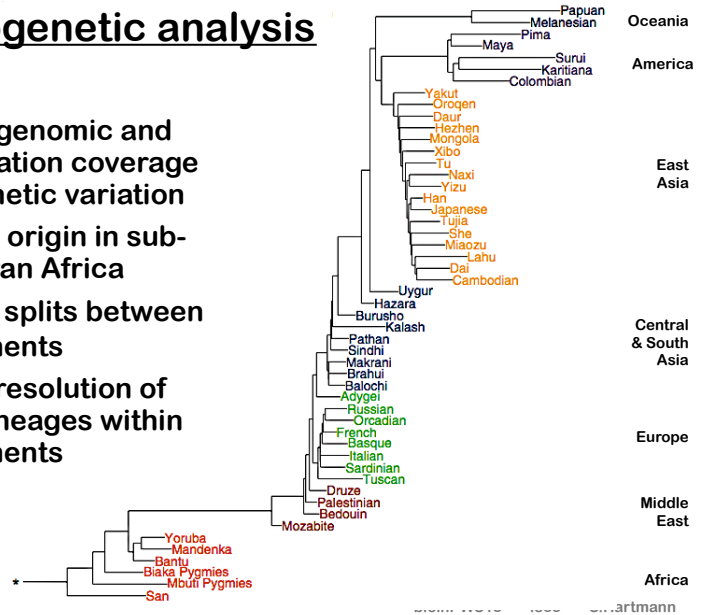






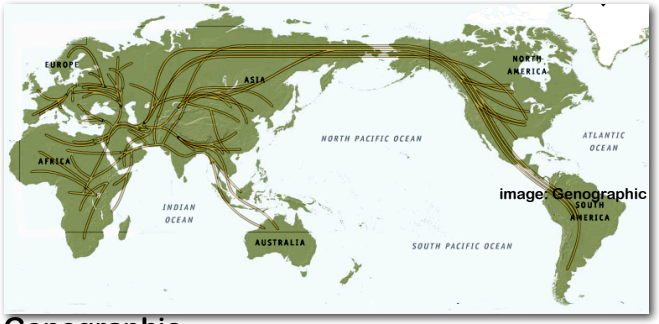
Phylogenetic analysis

- large genomic and population coverage of genetic variation
- single origin in sub-Saharan Africa
- major splits between continents
- good resolution of sub-lineages within continents



Peopling of the world

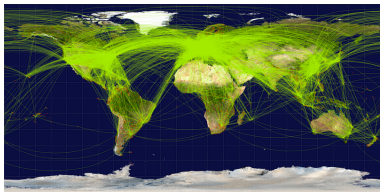
- consensus of general migration patterns & times
- open questions: timing, detailed routes, admixture, additional populations, demographics



Genographic

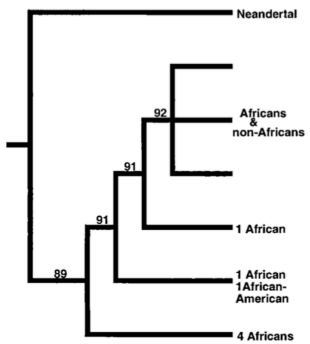
Genetic structure today ≠ yesterday

1. out of Africa
   - peopling of the world
2. migration, admixture, population replacement
   - African and Arab slave trades
   - European colonization of The Americas, Australia
   - European population movements after WWII
   - ...
3. today's globalization



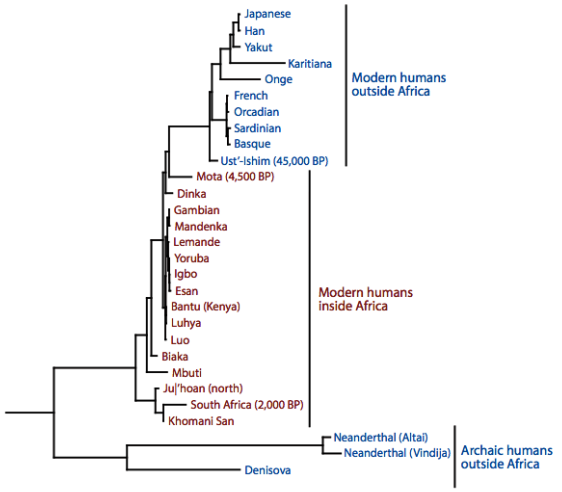
Neanderthal mt sequence data

mitochondrial HVR1 region (1997)
- Feldhofer cave, Germany
- sister lineage



additional mtDNA sequences,
from different locations
- Neanderthal sequences
  - were similar to each other
  - different from modern human sequences

Genome-wide analyses



P Skoglund & I Mathieson, 2018

Peopling of the Americas

migration during the last ice age

- Bering Land Bridge
- Native Americans / Paleo-Indians

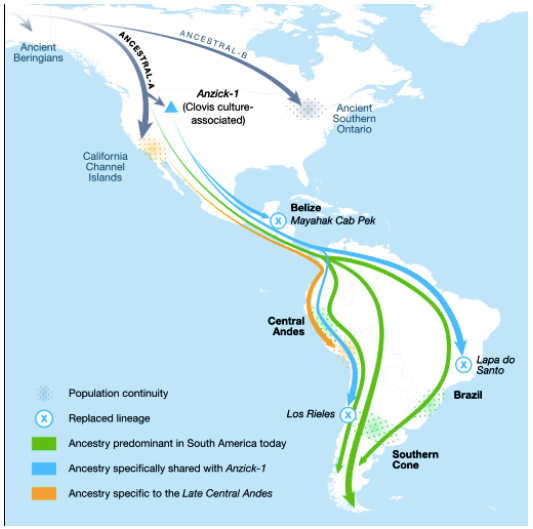
2010 census, USA:

- 0.9% only American Indian or Alaska Native
- most of European & African ancestry, or admixed

Peopling of the Americas

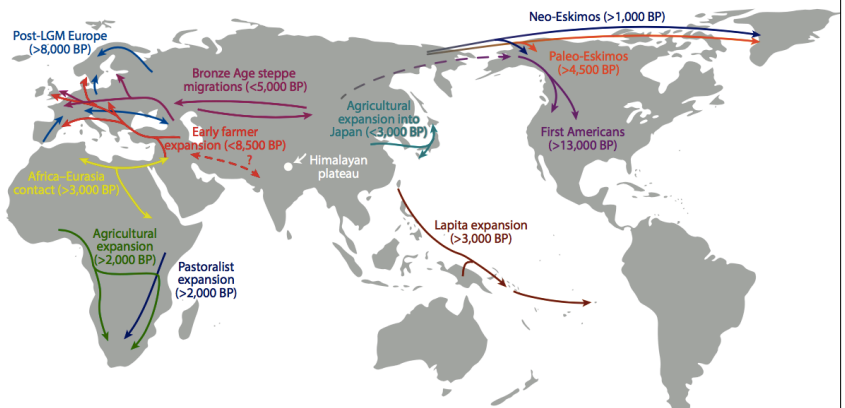
genome-wide analysis of 49 Central and South Americans up to 11,000 years old

Cell, 2018



Genome-wide analyses

population movements & expansions, based on aDNA studies of prehistoric modern humans



P Skoglund & I Mathieson, 2018

Human genetic data: applications

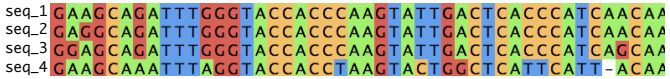
	selected genes	mitochondrial genomes	one (contemporary) nuclear genome	many contemporary nuclear genomes	contemporary and extinct nuclear genomes
basic research					
health & disease					
population structure & history					
adaptations					

Key terms and concepts

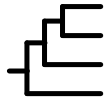
- genome, coverage, haplogroup, metadata
- SNPs, founder effect
- Sanger vs NGS: costs, throughput, read lengths
- relative cost of genome projects
- de novo assembly vs. mapping
  - principles, when best to use each, challenges
- one vs. many genomes from one species
  - general: opportunities?
  - applications: human genomes
    - human migrations: main findings

Today's exercise

- mitochondrial sequence data from 5 extant humans, 1 Neandertal, 1 chimpanzee
- compute & inspect an alignment



- compute, display, inspect relationships between sequences (maternal lineages)

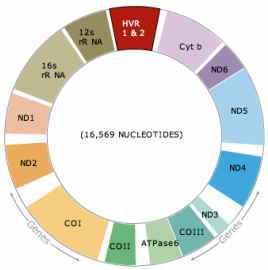


Today's exercise

- mitochondrial sequence data from 5 extant humans, 1 Neandertal, 1 chimpanzee

COX_sequences.fa
- cytochrome c oxidase subunit I
- 1562 bases

HVR_sequences.fa
- hypervariable region I
- 385 bases in length



Today's exercise

1. compute the alignment
   - EITHER locally, in the terminal, using ClustalW (just like two weeks ago)
   - OR online, using ClustalO, then download!
2. open JalView
   - inspect alignment
   - compute & inspect NJ phylogeny

do once for COX squences and once for HVR sequences