GENOME DATA ANALYSIS

LVA-Nr. 320.301 and 320.304

Irene Tiemann-Boege
Theresa Schwarz

JOHANNES KEPLER
UNIVERSITÄT LINZ

GENOMIC DATA ANALYSIS

Tuesday	8:30-11:00
Thursday	8:30-15:00
Friday	8:30-15:00

Instructor:

- Irene Tiemann-Boege/Theresa Schwarz
  Gruberstrasse. 40-44, 4020, Linz, Austria
  Tel: ++43 732 2468 7620
  Irene.Tiemann@jku.at

GOALS

- Bring genomics into the classroom
- To provide an introduction to genomic databases
- Focus on the analysis of DNA and proteins
- Apply concepts to solve basic exercises
- Combine theory and practice to help students solving common research problems in biology with the resources and information available in different online databases

JKU

FORMAT OF THE COURSE: LECTURE/COMPUTER LABS

- Lecture: Theory, concepts, examples of application of genomics, and case studies
- Computer lab: short lecture/exercises solved individually or in a group
- Computer labs will be based on solving questions of topics discussed in class
  - ☐ You will need to go to websites, use databases, and use software.
  - ☐ A written report is due at the end of each lab session

JKU

OUTLINE FOR THE COURSE—PART 1

- 1. The Human Genome Project Day 1
- 2. Genomic variation Day 2
- 3. Genome projects/ Comparative genomics Day 3
- 4. Emerging sequencing technologies Day 4
- 5. How genomics impacts our society Day 5

JKU

OUTLINE FOR THE COURSE—PART 2

- 1. Accessing information about DNA/proteins Day 1
- 2. SNP databases and PCR Day 2
- 3. Sequence alignments/BLAST Day 3
- 4. Protein analysis Day 4
- 5. Applications/ case studies Day 5

JKU

RESOURCES

- pdfs of the lectures available in Moodle
- E-learning: digital recordings of the lectures (accessible through moodle)
- Access to moodle:
- https://moodle.jku.at/

JKU

TEXTBOOK

- Some recommended textbooks:
- Bioinformatics and Functional Genomics by
  - Jonathan Pevsner (Wiley-Blackwell, 3rd edition 2015).
- A Primer of Genome Science by Greg Gibson (Spencer V. Muse
  - Publisher: Sinauer Associates, 3rd Edition 2008)
- Web-based resources

JKU

GRADING

- 50%: Final exam-You must pass the written exam in order to pass the course (≥60%)
  - ☐ short answer / multiple choice based on material of the lectures and the lab
  - ☐ Closed book, no computer.

Date of final exam:

☐ Thursday 07.06.2018 9:15 – 11:00 BA 9910

☐ Thursday 07.06.2018 9:15 – 11:00 K033C

JKU

GRADING

- 50% Final exam
- 10%: Small assay (answer 2 questions 1-2 pages)
- 35%: Labwork--Results from computer lab (no excuses for not handing in the report!)
- 5%: Attendance/Participation in class
- Grade scale:
- 1 90-100; 2 89-80; 3 70-79; 4 60-69; 5 0-59

JKU

OUTLINE FOR TODAY—PART 1

- Definition of bioinformatics/genomics
- The Human Genome Project-the start of genomics
- Sequencing the human genome
- Assembly: paired-end and shotgun sequencing
- Main conclusions of the human genome

JKU

WHAT IS GENOMICS?

- Study of structure, content and evolution of genomes—the DNA in our cells
- -sequencing of genomes
- -expression and function of genomes
- -evolution of the DNA
- -architecture of the genome
- -big databases publicly available online
- Why do we care to study the DNA in our cells?

JKU

WHY LOOK AT DNA?

- Proteins are much better predictors of phenotype
- Why not study the proteins?
  - Sequencing DNA is technically much easier than sequencing aminoacids
- Why study the DNA in our cells?
  - Central dogma of transmission of genetic information

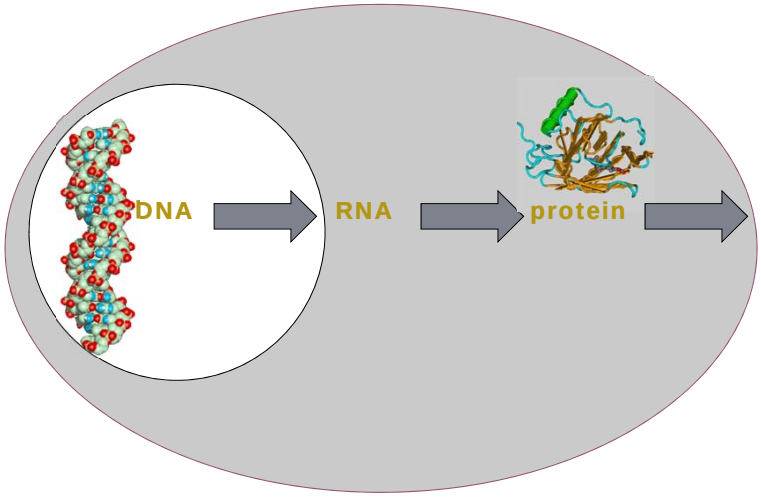
JKU

WHY STUDY GENOMES?

- The genome (DNA) gives us information about the transcriptome and proteome
- Available technology to obtain the DNA sequence (before sequencing was discovered scientist compared proteins-allozymes)
- DNA is easy to access- available and stable even in fossils

JKU

Central dogma of molecular biology

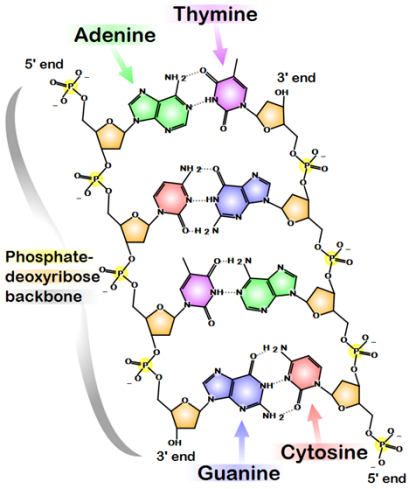


genome → transcriptome → proteome

Central dogma of bioinformatics and genomics

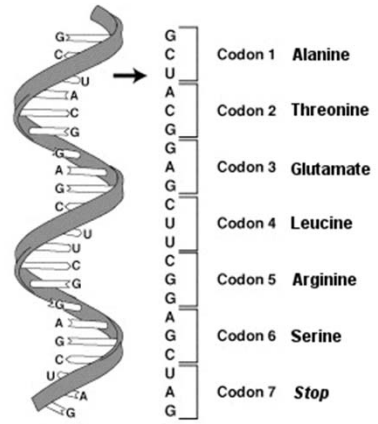
15

GENETIC INFORMATION IS CONTAINED IN THE DNA



JKU

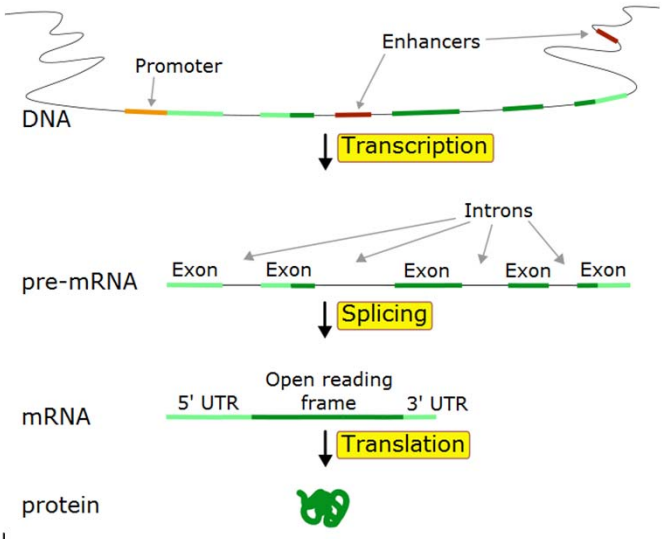
THE GENETIC CODE



1st base	2nd base T		2nd base C		2nd base A		2nd base G	
T	TTT	(Phe/F) Phenylalanine	TCT	(Ser/S) Serine	TAT	(Tyr/Y) Tyrosine	TGT	(Cys/C) Cysteine
	TTC	(Phe/F) Phenylalanine	TCC	(Ser/S) Serine	TAC	(Tyr/Y) Tyrosine	TGC	(Cys/C) Cysteine
	TTA	(Leu/L) Leucine	TCA	(Ser/S) Serine	TAA	Stop (*Ochre*)	TGA	Stop (*Opal*)
	TTG	(Leu/L) Leucine	TCG	(Ser/S) Serine	TAG	Stop (*Amber*)	TGG	(Trp/W) Tryptophan
C	CTT	(Leu/L) Leucine	CCT	(Pro/P) Proline	CAT	(His/H) Histidine	CGT	(Arg/R) Arginine
	CTC	(Leu/L) Leucine	CCC	(Pro/P) Proline	CAC	(His/H) Histidine	CGC	(Arg/R) Arginine
	CTA	(Leu/L) Leucine	CCA	(Pro/P) Proline	CAA	(Gln/Q) Glutamine	CGA	(Arg/R) Arginine
	CTG	(Leu/L) Leucine	CCG	(Pro/P) Proline	CAG	(Gln/Q) Glutamine	CGG	(Arg/R) Arginine
A	ATT	(Ile/I) Isoleucine	ACT	(Thr/T) Threonine	AAT	(Asn/N) Asparagine	AGT	(Ser/S) Serine
	ATC	(Ile/I) Isoleucine	ACC	(Thr/T) Threonine	AAC	(Asn/N) Asparagine	AGC	(Ser/S) Serine
	ATA	(Ile/I) Isoleucine	ACA	(Thr/T) Threonine	AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine
	ATG[n]	(Met/M) Methionine	ACG	(Thr/T) Threonine	AAG	(Lys/K) Lysine	AGG	(Arg/R) Arginine
G	GTT	(Val/V) Valine	GCT	(Ala/A) Alanine	GAT	(Asp/D) Aspartic acid	GGT	(Gly/G) Glycine
	GTC	(Val/V) Valine	GCC	(Ala/A) Alanine	GAC	(Asp/D) Aspartic acid	GGC	(Gly/G) Glycine
	GTA	(Val/V) Valine	GCA	(Ala/A) Alanine	GAA	(Glu/E) Glutamic acid	GGA	(Gly/G) Glycine
	GTG	(Val/V) Valine	GCG	(Ala/A) Alanine	GAG	(Glu/E) Glutamic acid	GGG	(Gly/G) Glycine

JKU

DNA CODES FOR PROTEINS



JKU

DOES DNA ONLY CODE FOR PROTEINS?

- Genomic DNA → protein coding DNA→ mRNA
- Non-Protein Coding DNA→ rRNA, tRNA or iRNA
- The world of the RNA and its importance in gene regulation

JKU

GENOMICS VS. BIOINFORMATICS

- Interface of biology and computers
- Analysis of proteins, genes and genomes using computer algorithms and computer databases
- The tools of bioinformatics are used to make
- sense of the billions of base pairs of DNA that are sequenced by genomics projects.

JKU

GENOMICS VS. BIOINFORMATICS

- Genomics is different from bioinformatics
  - Genomics is the analysis of genomes,
  - Involves also experimental approaches (e.g sequencing)
  - Technological developments
  - Curation of sequences
  - Functional sequence characterization
  - Comparative genomics

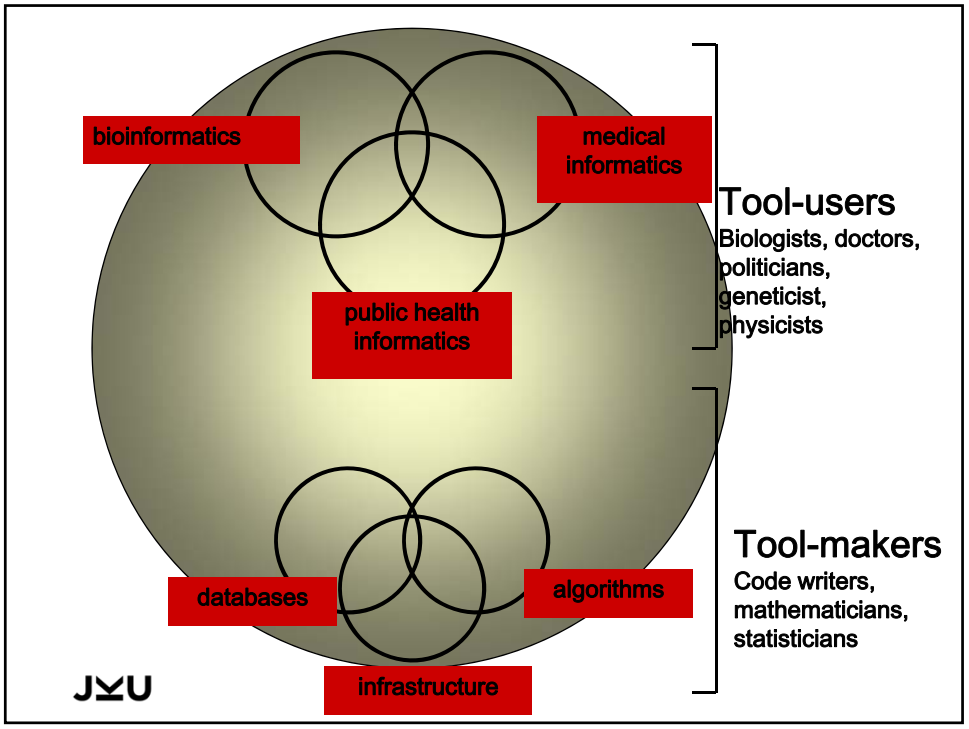
JKU

WHAT TOOLS DO WE NEED FOR GENOMICS?

- Biological questions can be approached from different levels
- single genes and proteins
- cellular pathways and networks
- whole genomic responses

- To study thousands of genes and proteins we need: sequencing technology, computers, mathematical algorithms, and access to servers (databases)—internet!

JKU



OUTLINE FOR TODAY—PART 1

- Definition of bioinformatics/genomics
- The Human Genome Project-the start of genomics
- Sequencing the human genome
- Assembly: paired-end and shotgun sequencing
- Main conclusions of the human genome

JKU

HUMAN GENOME PROJECT

- Sequencing the human genome.
- How big is our haploid genome?
  - $3x10^9$ nucleotides
- Facilitate molecular biological research (genetics, systems biology, medicine)
- Historically, create a fine genetic map of our genome

JKU

HISTORY OF THE HUMAN GENOME

- Two efforts:
  - Publicly funded (lead by Francis Collins and Human Genome Consortia)
    - Started in 1993—outlook: 5 years to finish the first draft
    - High resolution genetic mapping
  - Private effort (lead by TIGR/Celera by Craig Venter)
    - Commercialize information as patents, access rights, etc.

JKU

HISTORY OF THE HUMAN GENOME

- Public effort
  - Lead by multiple countries (US, Germany, UK, France, Japan, China, Canada, India)
  - Leader: Francis Collins
  - Cost: $3-billion
  - Hierarchical sequencing—sequence step-by step.
- Private effort
  - Lead by Celera (before: TIGR institute, today: Craig Venter Institute)
  - Leader Craig Venter
  - Cost: $0.3 billion
  - Shotgun sequencing—sequence and assemble random fragments

JKU

WHO WAS SEQUENCED?

Public effort

- Genomes of a ethnically diverse panel of donors
- >50 donors
- first draft: male donors

- Construct PAC/BAC libraries

Private effort

- Genomes of a ethnically diverse panel of donors
- >21 donors
- 2 male; 3 female donors (1 African, 1 Asian, 1 Hispanic, 2 Caucasian)

- Permanent cell lines
- 2-, 10-, 50- kb libraries

JKU

Table 3 Total human sequence deposited in the HTGS division of GenBank

Sequencing centre	Total human sequence (kb)	Finished human sequence (kb)
Whitehead Institute, Center for Genome Research*	1,196,888	46,560
The Sanger Centre*	970,789	284,353
Washington University Genome Sequencing Center*	765,898	175,279
US DOE Joint Genome Institute	377,998	78,486
Baylor College of Medicine Human Genome Sequencing Center	345,125	53,418
RIKEN Genomic Sciences Center	203,166	16,971
Genoscope	85,995	48,808
GTC Sequencing Center	71,357	7,014
Department of Genome Analysis, Institute of Molecular Biotechnology	49,865	17,788
Beijing Genomics Institute/Human Genome Center	42,865	6,297
Multimegabase Sequencing Center; Institute for Systems Biology	31,241	9,676
Stanford Genome Technology Center	29,728	3,530
The Stanford Human Genome Center and Department of Genetics	28,162	9,121
University of Washington Genome Center	24,115	14,692
Keio University	17,364	13,058
University of Texas Southwestern Medical Center at Dallas	11,670	7,028
University of Oklahoma Advanced Center for Genome Technology	10,071	9,155
Max Planck Institute for Molecular Genetics	7,650	2,940
GBF – German Research Centre for Biotechnology	4,639	2,338
Cold Spring Harbor Laboratory Lita Annenberg Hazen Genome Center	4,338	2,104
Other	59,574	35,911
Total	4,338,224	842,027

Total human sequence deposited in GenBank by members of the International Human Genome Sequencing Consortium, as of 8 October 2000. The amount of total sequence (finished plus draft plus predraft) is shown in the second column and the amount of finished sequence is shown in the third column. Total sequence differs from totals in Tables 1 and 2 because of inclusion of padding characters and of some clones not used in assembly. HTGS, high throughput genome sequence.

*These three centres produced an additional 2.4 Gb of raw plasmid paired-end reads (see Table 4), consisting of 0.99 Gb from Whitehead Institute, 0.66 Gb from The Sanger Centre and 0.75 Gb from Washington University.

WHAT ADVANCES MADE THIS TASK POSSIBLE:

- Automation—creation of clones
- Capillary sequencers with automatic base calling

Sequencing production line at the Whitehead Institute (MIT).

"The system consists of custom-designed factory-style conveyor belt robots that perform all functions from purifying DNA from bacterial cultures through setting up and purifying sequencing reactions." IHGSC. 2001. Nature.

JKU

IHGSC—INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM

- Publicly funded effort (objectives)
  - High resolution genetic and physical maps
  - Attain a complete sequence by 2005 (finished in 2001); curation finished: ~2004 (with 1 error/10,000bp)
  - Identification of all genes
    - Identify open reading frames
    - Construct expression sequence tags (What are EST?)
    - Comparative and functional genomics (compare sequence with other organisms)
  - Single nucleotide polymorphism maps (HapMap)

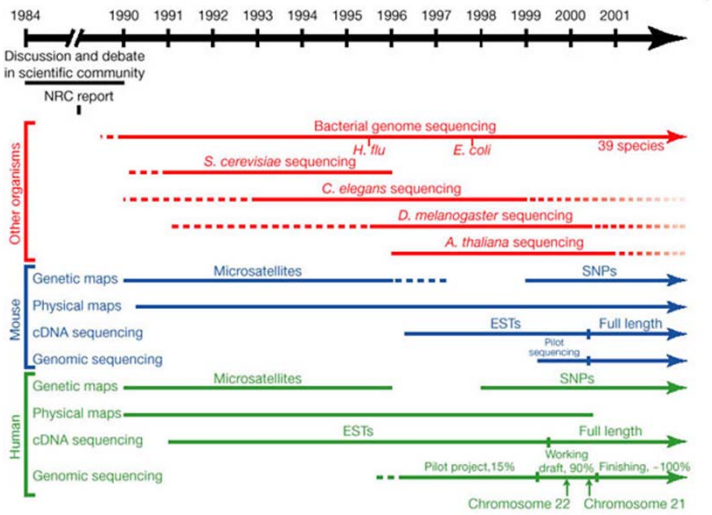
JKU

AREAS DEVELOPED IN PARALLEL WITH THE HUMAN GENOME PROJECT

- Bioinformatic tools
  - Gene prediction tools
  - Protein folding
  - Human genome browsers (UCSC, Ensembl, NCBI))
- DNA sequencing
  - automation of high throughput sequencing
  - sequencing capacity increases exponentially every year since 1998
- Ethical, Legal, and Social Issues
  - Non-discrimination act based on genetic information
  - Understand human genetic variation
- Model Organisms
  - Mouse, fly, worm, yeast

JKU

THE HUMAN GENOME WAS NOT THE FIRST TO BE SEQUENCED



Taken from IHGSC. 2001. Nature.

JKU

OUTLINE FOR TODAY—PART 1

- Definition of bioinformatics/genomics
- The Human Genome Project-the start of genomics
- Sequencing the human genome
- Assembly: paired-end and shotgun sequencing
- Main conclusions of the human genome

JKU

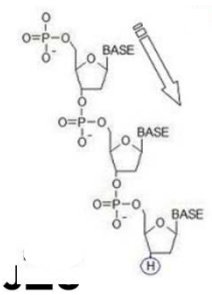
SANGER SEQUENCING

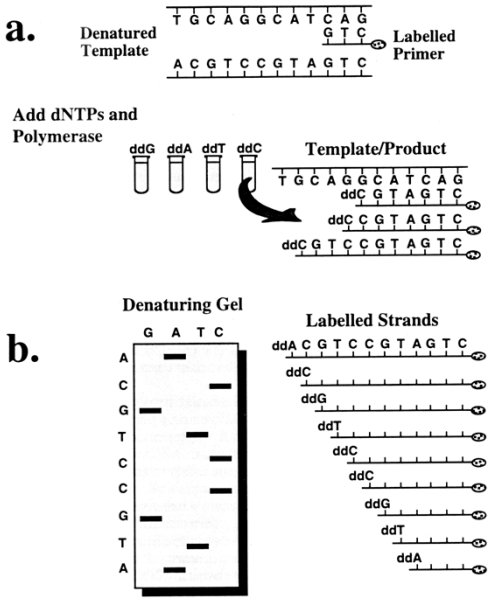
- Sanger sequencing is an extension product
- A single extension product is not sufficient to be visualized on a gel
- Before sequencing the genomic DNA needs to be amplified to produce enough copies that are part of a discrete fluorescent band
- Methods of amplification:
  - ☐ PCR
  - ☐ Cloning

JKU

SANGER SEQUENCING

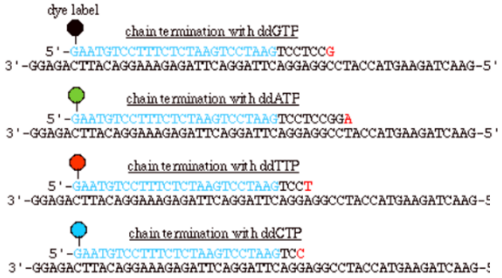
- What do we need?
  - ☐ Primers, dNTPs, ddNTPs, Polymerase, genomic DNA

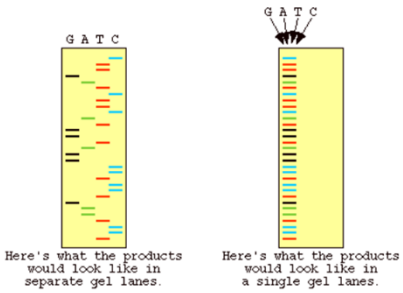




AUTOMATED SANGER SEQUENCING

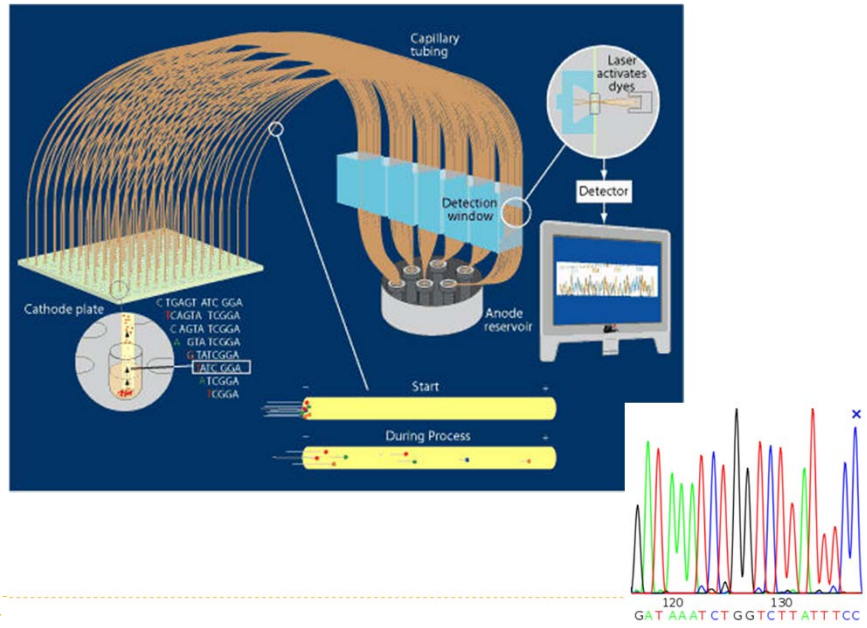
- Capillary electrophoresis





JKU

Automated Sanger sequencing: capillary electrophoresis



OUTLINE FOR TODAY—PART 1

- Definition of bioinformatics/genomics
- The Human Genome Project-the start of genomics
- Sequencing the human genome
- Assembly: paired-end and shotgun sequencing
- Main conclusions of the human genome

JKU

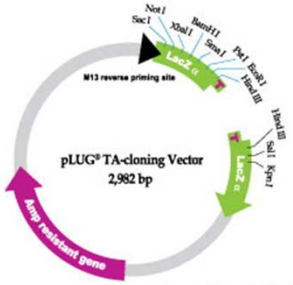
STRATEGY TO SEQUENCE AN UNKNOWN DNA

- What primers should be used?
- How can the primers be designed if the annealing sequence is not known yet?
- Benefits of cloning

JKU

CLONING FOLLOWED BY SEQUENCING

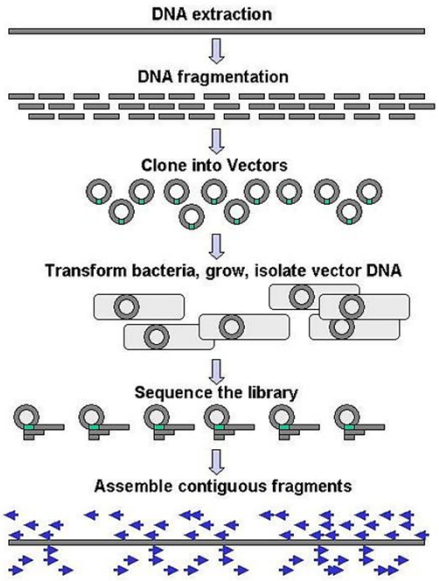
- Random pieces of genomic DNA are inserted into a vector, then introduced into a bacteria (E. coli) and amplified via the bacterial replication machinery--cloning.
- Vector sequences (M13 reverse and forward priming site) are used for sequencing



Element	Position (bp)
LacZα fragment	216 ~ 643
LacZα start cordon	216
M13 reverse priming site	204 ~ 220
M13 forward priming site	439 ~ 454
Multiple cloning site	272 ~ 404
Ampicillin resistance gene	1,188 ~ 2,308
f1 origin of replication	594 ~ 1,050
ColE1 origin of replication	2,042 ~ 2,082

Figure 3. Map of pLUG® TA-cloning vector

JKU



WHAT IS PAIR-END SEQUENCING (ALSO KNOWN AS SHOTGUN SEQUENCING)?

- None of the reads cover the full sequence
- By several rounds of fragmentation and sequencing the original sequence can be reconstructed

Strand	Sequence
Original	AGCATGCTGCAGTCATGCTTAGGCTA
First shotgun sequence	AGCATGCTGCAGTCATGCT------- -------------------TAGGCTA
Second shotgun sequence	AGCATG-------------------- ------CTGCAGTCATGCTTAGGCTA
Reconstruction	AGCATGCTGCAGTCATGCTTAGGCTA

JKU

PAIR-END SEQUENCING

- Useful information to reconstruct the sequence from pairwise end sequencing
  - Pairwise end sequencing: sequences obtained from two directions
  - M13 Forward and M13 reverse primers complementary to the vector sequence
  - Orientation of two sequences relative to each other (inversions)
  - Space between two sequences

JKU

HOW CAN A SEQUENCE BE RECONSTRUCTED BY SEQUENCING ONLY PAIRED ENDS?

- The concept of coverage
  - Pair ends will rarely overlap (2 kb clone vs. 500bp sequence reads)
  - Except if multiple rounds of fragmentation and sequencing are used
  - Coverage: the average number of reads representing a given nucleotide in the reconstructed sequence

JKU

ESTIMATION OF COVERAGE

- Coverage = NL/G
  - N= No. reads
  - L= Ave. read length
  - G = genome size

  - Eg. Estimate the coverage of a 4000bp genome, with 10 reads, 400bp in length each?

JKU

COVERAGE OF THE HUMAN GENOME

- IHGSC—5x coverage
- Celera—12x coverage
- The higher the coverage, the less the errors in base calling and assembly (especially for shot-gun sequencing)
- Still some gaps for 1% of euchromatic human genome

JKU

SEQUENCING BACS

1. A piece of the organism's DNA is inserted in BACs (~150kb)

2. BACs are a functional fertility plasmid (or F-plasmid) that are used for cloning in *E. coli*

1. F-plasmids contain partition genes that promote the even distribution of plasmids after bacterial cell division

2. The sequenced parts of the BACs are assembled *in silico* resulting in a fragment of the genomic sequence of the organism.

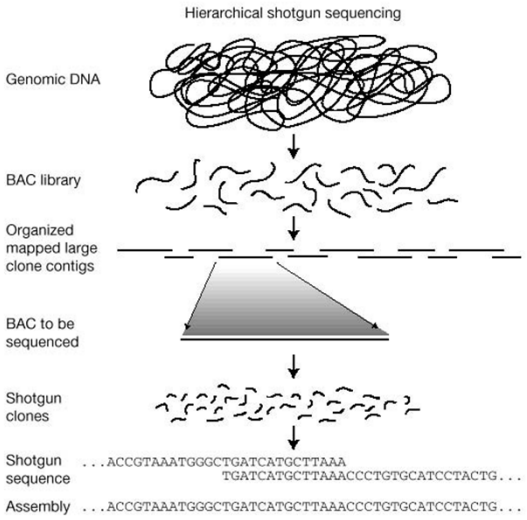
3. BACs are replaced with faster and less laborious sequencing methods like whole genome shotgun sequencing and now more recently next-generation sequencing.

JKU

SEQUENCING APPROACH TAKEN BY IHGSC- FIRST CREATE A ROADMAP (BACS)

1. A library is constructed by fragmenting the target genome and cloning it into a large-fragment cloning vector (BAC).
2. BACS are ~150kb in size
3. Each BAC is organized into a physical map.
4. Individual BAC clones are selected and sequenced by the random shotgun strategy.
5. The clone sequences are assembled to reconstruct the sequence of the genome.

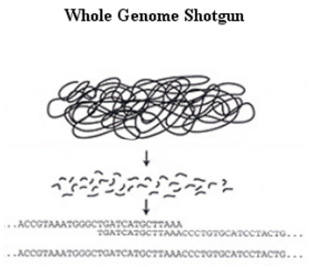
Taken from IHGSC. 2001. Nature.



JKU

CAN A COMPLETE GENOME BE ASSEMBLED ONLY BY SHOTGUN SEQUENCING?

- The Celera approach-shotgun sequencing
  - Shear genome into 2-, 10-, 50kb fragments
  - Insert fragments in clones
  - Sequence only clone ends: 200-1000bp—same reaction mix can be used for all the clones—PAIR-END SEQUENCING
  - Need overlapping reads—requires several rounds of fragmentation and sequencing



JKU

CAN A COMPLETE GENOME BE ASSEMBLED ONLY BY SHOTGUN SEQUENCING?

- Craig Venter is crazy!
  - Against: ability to correctly link these regions (eg. repeating regions)
  - For: It is possible to sequence the whole genome at once with shotgun sequencing using large arrays of sequencers
  - Sequence assembly programs are considerably improved (but sequence road map provided by the IHGSC was very useful)
  - computing power becomes cheaper and faster

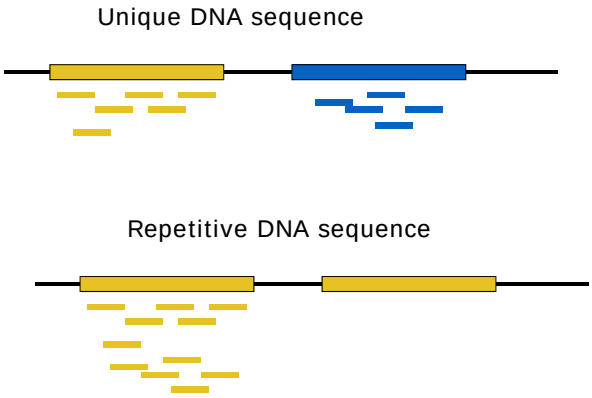
JKU

HOW MANY SEQUENCE READS ARE REQUIRED FOR SEQUENCING THE HUMAN GENOME AT 12X COVERAGE?

- Size of human genome: $3x10^9$bp
- No. of bp with 12x oversampling = $3.6x10^{10}$bp
- Length of sequence read: 543bp
- N= $7.2x10^7$ reads; Celera produced $3x10^7$ reads
- To assemble $7.2x10^7$ reads:
  - $(7.2x10^7)^7 = 10^{50}$ comparisons have to be made

JKU

CAN SHOTGUN SEQUENCING BE USED FOR REPETITIVE REGIONS?



JKU

OUTLINE FOR TODAY—PART 1

- Definition of bioinformatics/genomics
- The Human Genome Project-the start of genomics
- Sequencing the human genome
- Assembly: paired-end and shotgun sequencing
- Main conclusions of the human genome

JKU

JUNE 2000—ANNOUNCING THE COMPLETION OF THE HUMAN GENOME

- 2003—complete genome is published
- 2006—last sequenced chromosome
- Parts of heterochromatic areas un-sequenced (centromeres, telomeres), mulitgene families, some gaps (8%)
- Every year, the human genome is updated and a new version is released

JKU

Our changing genome

Summary

Assembly:	GRCh37.p12, Feb 2009
Database version:	73.37
Base Pairs:	3,324,592,091
Golden Path Length:	3,101,804,739

Gene counts (Primary assembly)

Coding genes:	20,769
Short Non coding genes:	9,079
Long Non coding genes:	13,564
Pseudogenes:	14,165
Gene transcripts:	195,565

Other

Genscan gene predictions:	48,461
Short Variants (SNPs, indels, somatic mutations):	55,286,608
Structural variants:	10,343,072

Summary

Assembly	GRCh38.p2 (Genome Reference Consortium Human Build 38), INSDC Assembly GCA_000001405.17, Dec 2013
Database version	79.38
Base Pairs	3,384,269,757
Golden Path Length	3,096,649,726

Gene counts (Primary assembly)

Coding genes	20,300 (incl 519 readthrough)
Non coding genes	24,885
Small non coding genes	7,715
Long non coding genes	14,863 (incl 193 readthrough)
Misc non coding genes	2,307
Pseudogenes	14,424 (incl 4 readthrough)
Gene transcripts	198,622

Other

Genscan gene predictions	50,421
Short Variants	65,897,584
Structural variants	4,168,103

Our changing genome
http://www.ensembl.org/Homo_sapiens/Info/Annotation

MAIN CONCLUSIONS OF HUMAN GENOME PROJECT

- 1. We have about the same number of genes as fish and plants, and not that many more genes than worms and flies.
  - Fugu rubripes (pufferfish): 31,000 to 38,000 (2009 estimate: ~19,000)
  - Arabidopsis thaliana (thale cress): 26,000
  - Caenorhabditis elegans (worm): 19,000
  - Drosophila melanogaster (fly): 13,000

JKU

MAIN CONCLUSIONS OF HUMAN GENOME PROJECT

- The human proteome is far more complex than the set of proteins encoded by invertebrate genomes.
- Vertebrates have a more complex mixture of protein domain architectures.
- The human genome displays greater complexity in its processing of mRNA transcripts by alternative splicing.
- 98% of the genome does not code for genes
  - >50% of the genome consists of repetitive DNA
  - ~25,000 non-coding genes

JKU

Non-Coding DNA (ncRNA)

- Not all the DNA is coding (Coding DNA is 1% of our genome)
  - Only 26% of genes code for proteins
  - Most DNA is transcribed into ncRNA
    - transfer RNAs and ribosomal RNA
  - ~1500 microRNAs (miRNA; 22nts)
  - The function of the majority of ncRNA is not known
    - control of chromosome dynamics, splicing, translational inhibition, and mRNA destruction
- our protein-centric view of gene expression and regulation is undergoing a change.

JKU

PROPERTIES OF HUMAN GENES

- Genes vary in size

TABLE 2.1: PROPERTIES OF HUMAN GENES

	Mean values for 1804 genes	Dystrophin gene (*DMD*)	Sex-determining region, Y gene (*SRY*)
Exon length	145 bp	180 bp mean	612 bp
Exon number	8.8	79	1
Intron length	3365 bp	30,000 bp mean	–
5′ UTR length	300 bp	200 bp	140 bp
3′ UTR length	770 bp		133 bp
Coding sequence length	1340 bp (447 aa)	14,000 bp	612 bp (204 aa)
Genomic extent	27 kb	2700 kb	1 kb

[Data from International Human Genome Sequencing Consortium (2001) *Nature* 409, 860; Roberts RG et al. (1993) *Genomics* 16, 536; Koenig M et al. (1987) *Cell* 50, 509; Behlke MA et al. (1993) *Genomics* 17, 736.]
aa, amino acid; UTR, untranslated region.

Table 2.1 Human Evolutionary Genetics, 2nd ed. (© Garland Science 2014)

JKU

BROAD GENOMIC LANDSCAPE: GC CONTENT

- The overall GC content of the human genome is 41%
- Some genomic regions are GC-rich, while some are GC-poor
- Isochores: large DNA segments (e.g. >300 kb) with similar GC content

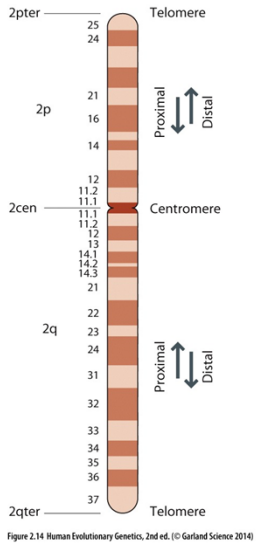


Figure 2.14 Human Evolutionary Genetics, 2nd ed. (© Garland Science 2014)

JKU

BROAD GENOMIC LANDSCAPE: CPG ISLANDS

- Dinucleotides of CpG are under-represented
- CpG dinucleotides are often methylated
- Methylated CpGs associate with control of gene expression, gene silencing, genomic imprinting, and X-chromosome inactivation.

JKU

JOHANNES KEPLER
UNIVERSITÄT LINZ

QUESTIONS?

JOHANNES KEPLER
UNIVERSITÄT LINZ
Altenberger Straße 69
4040 Linz, Österreich
www.jku.at