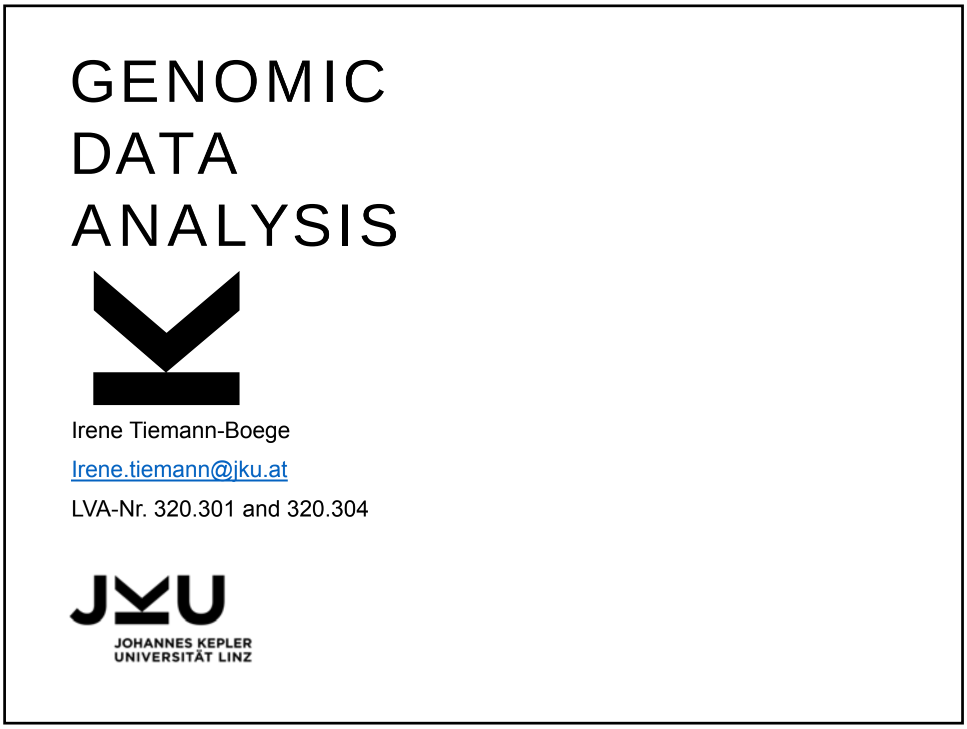
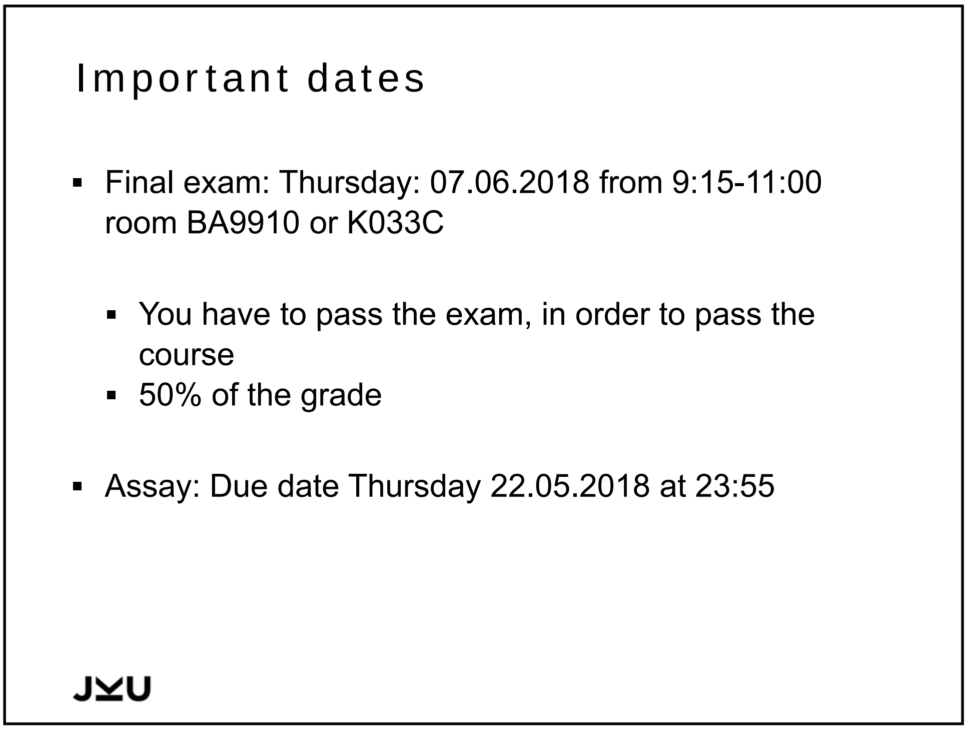


GENOMIC DATA ANALYSIS

Irene Tiemann-Boege

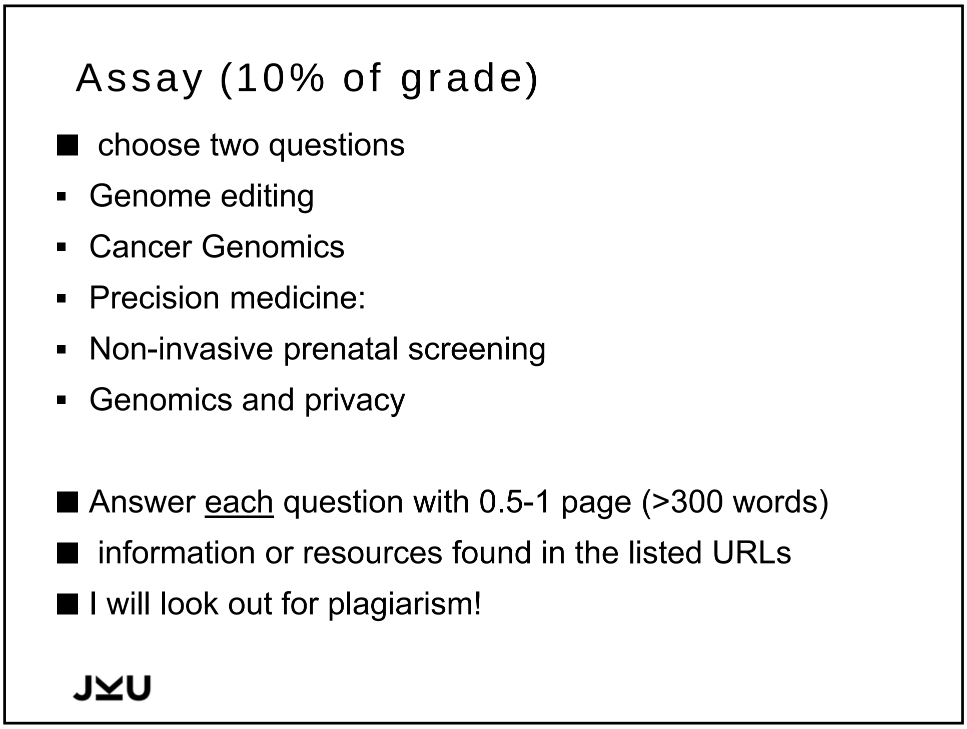
Irene.tiemann@jku.at

LVA-Nr. 320.301 and 320.304



Important dates

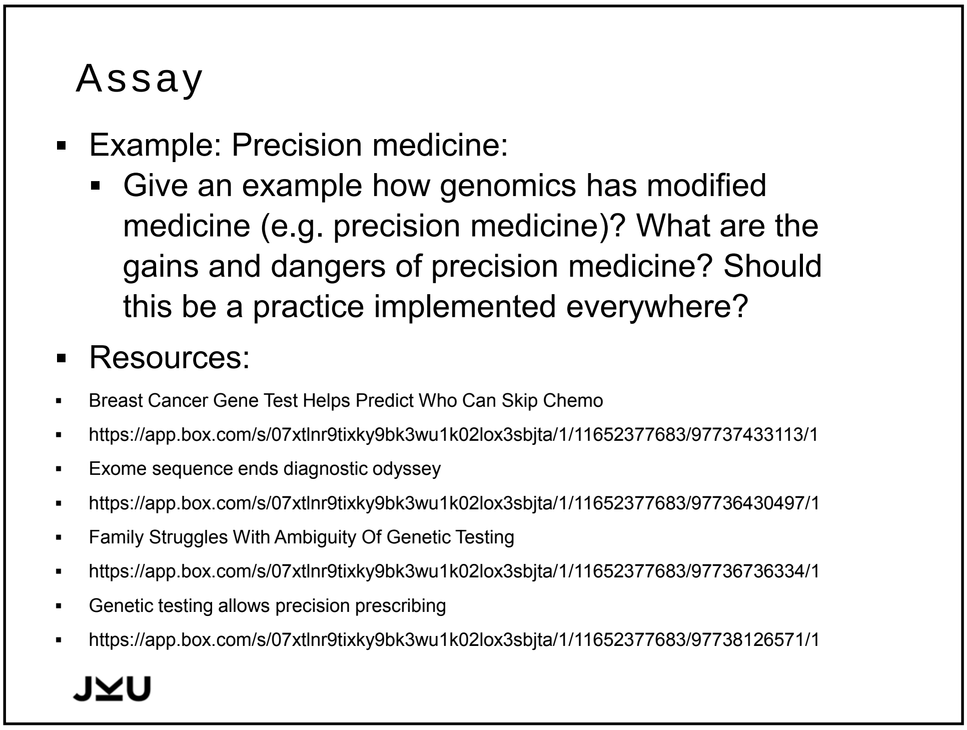
- Final exam: Thursday: 07.06.2018 from 9:15-11:00 room BA9910 or K033C
  - You have to pass the exam, in order to pass the course
  - 50% of the grade
- Assay: Due date Thursday 22.05.2018 at 23:55



Assay (10% of grade)

- choose two questions
- Genome editing
- Cancer Genomics
- Precision medicine:
- Non-invasive prenatal screening
- Genomics and privacy

- Answer <u>each</u> question with 0.5-1 page (>300 words)
- information or resources found in the listed URLs
- I will look out for plagiarism!



Assay

- Example: Precision medicine:
  - Give an example how genomics has modified medicine (e.g. precision medicine)? What are the gains and dangers of precision medicine? Should this be a practice implemented everywhere?
- Resources:
- Breast Cancer Gene Test Helps Predict Who Can Skip Chemo
- https://app.box.com/s/07xtlnr9tixky9bk3wu1k02lox3sbjta/1/11652377683/97737433113/1
- Exome sequence ends diagnostic odyssey
- https://app.box.com/s/07xtlnr9tixky9bk3wu1k02lox3sbjta/1/11652377683/97736430497/1
- Family Struggles With Ambiguity Of Genetic Testing
- https://app.box.com/s/07xtlnr9tixky9bk3wu1k02lox3sbjta/1/11652377683/97736736334/1
- Genetic testing allows precision prescribing
- https://app.box.com/s/07xtlnr9tixky9bk3wu1k02lox3sbjta/1/11652377683/97738126571/1

Outline for today-part 1

- Methods to detect genomic variation
- Sequencing Projects
- Understanding a genome sequence
- Structural features of a genome

Methods to detect genomic variation

- variation in the genome exists over a wide range of physical scales, from single nucleotides to whole chromosomes-methods such also suit the scale

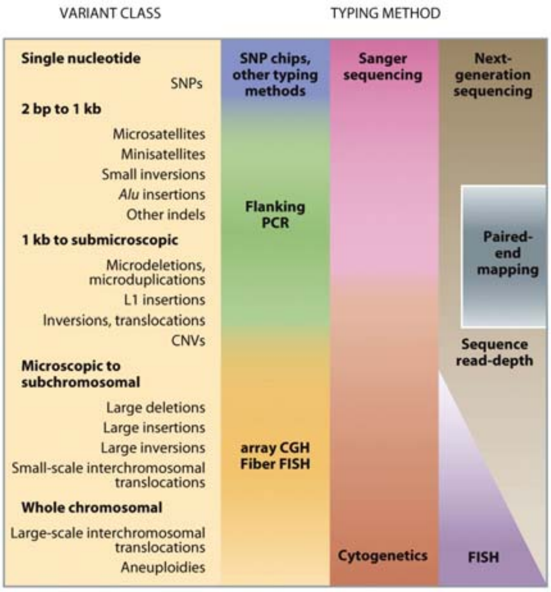


Figure 4.1 Human Evolutionary Genetics, 2nd ed. (© Garland Science 2014)

Sources of DNA

- Saliva or buccal (cheek) cells offer a less invasive alternative to blood.
- Semen are used in studies of haploid genomes or recombination and germline mutations
- Hair: the root contains much cellular material for DNA extraction. The hair shaft contains small amounts of very fragmented DNA since the root contains much cellular material for DNA extraction.
- Cell lines are established from primary sampled tissues, they can be cultured, sometimes indefinitely, yielding unlimited amounts of DNA
- Whole genome amplification can increase the amounts of DNA

JKU

SNP Genotyping—Low throughput

- PCR based-RFLP (restriction fragment length polymorphisms)
- A G/A SNP creates or destroys a BamHI restriction site.
- A second (constitutive) BamHI site within the same PCR amplicon provides a positive control for the completeness of restriction enzyme digestion.

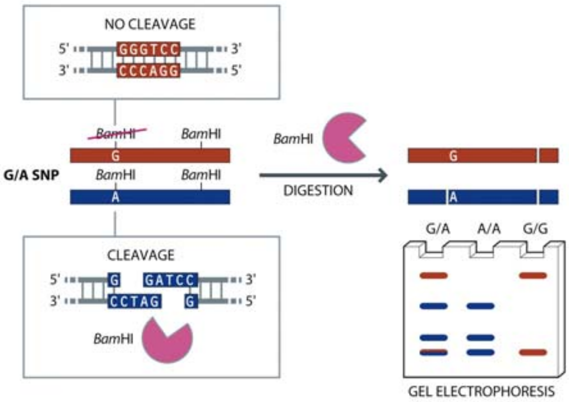
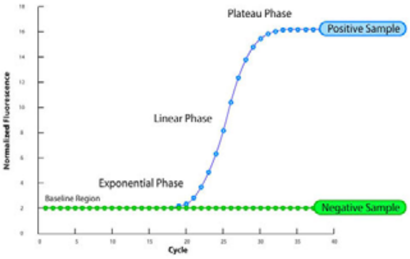


Figure 4.10 Human Evolutionary Genetics, 2nd ed. (© Garland Science 2014)

JKU

Medium throughput genotyping assays

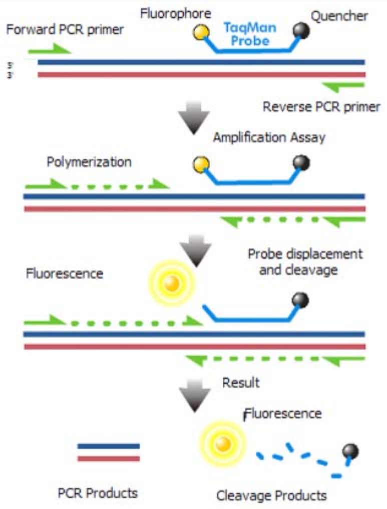
- Real time (qPCR): PCR is carried out in specialized thermocycler with camera/sensor that measures light and monitors the increase of PCR products with every cycle
- DNA amount is proportional to the fluorescence of intercalating dye (EvaGreen) hybridization probes (TaqMan probes)



https://www.youtube.com/watch?v=FIgGKkcLLuo

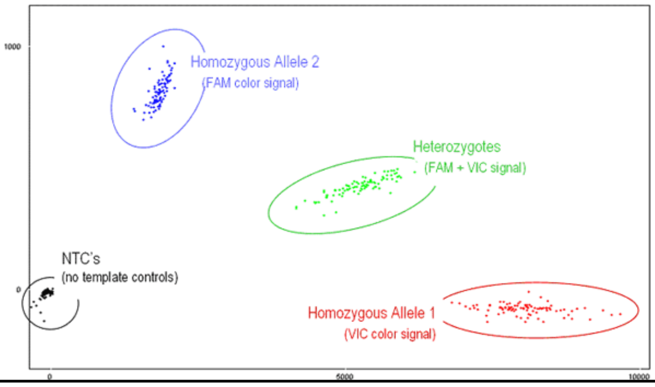
Medium throughput genotyping assays

- Taq Man-
  - Each allele hybridizes with a different labelled probe
  - Probes have a fluorophore at the 5’ end and a quencher at the 3’ end
  - Only perfectly hybridized probes are degraded releasing a fluorescent signal
  - https://www.youtube.com/watch?v=ob3teCrpgxY



TaqMan-real time PCR results

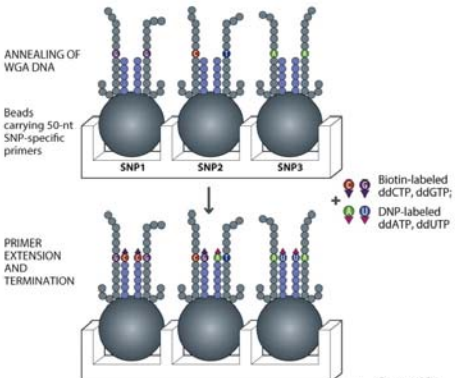
- Taq Man-
  - Each probe specific for an allele releases a fluorescent signal in a different color.
  - Genotyping by monitoring the increase of fluoresce with cycle number for each particular fluorphore
  - Genotyping by comparing the intensity fluorophore 1 (x axis) vs fluorphore 2 (y axis)



JKU

SNP Genotyping—High throughput

- throughput SNP chips simultaneously analyze more than 1 million SNPs
  - microarray-based technologies (SNP chips), 1 million SNPs 99% accuracy; ~$300
  - Infinium assay:
  - based on primer extensions
  - DNA fragments (amplified via whole genome amplification) are annealed to an array of silica beads (each with many copies of a 50-nt oligonucleotide with a 3' end adjacent to a particular SNP site)
  - Primer extension incorporates the complementar modified dideoxynucleotide, which terminates synthesis, and also allows detection via fluorescence.


ANNEALING OF WGA DNA
Beads carrying 50-nt SNP-specific primers
SNP1
SNP2
SNP3
+
Biotin-labeled ddCTP, ddGTP;
DNP-labeled ddATP, ddUTP
PRIMER EXTENSION AND TERMINATION

JKU

Infinium array

- Extension products are probed with fluorescent antibodies specific for the modified nucleotides.
- The colors are screened to call a sample homozygote or heterozygote
- Infinium assay:
  - Homozygous SNPs will produce single-color fluorescence, while heterozygous SNPs will produce approximately equal ratios of two colors, allowing the genotype at each bead to be called.

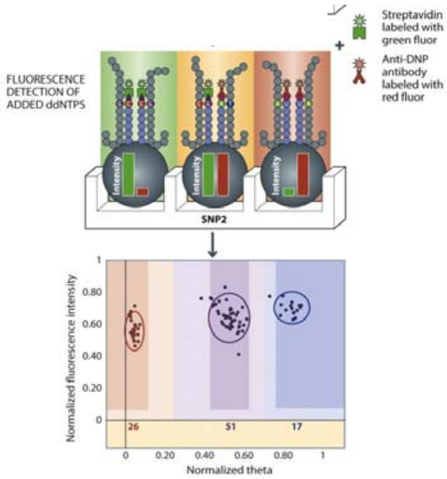
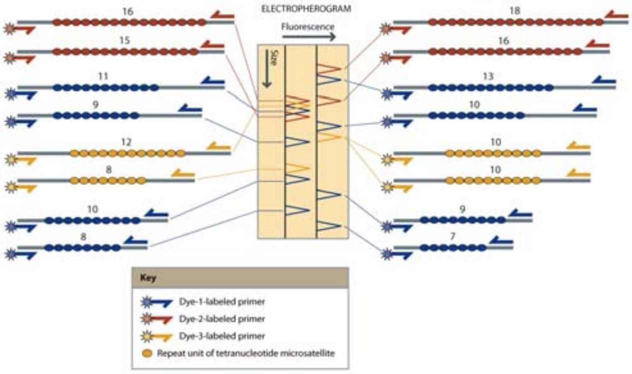

Streptavidin labeled with green fluor
+
Anti-DNP antibody labeled with red fluor
FLUORESCENCE DETECTION OF ADDED ddNTPS
Intensity
Intensity
Intensity
SNP2
Normalized fluorescence intensity
1
0.80
0.60
0.40
0.20
0
26
51
17
0
0.20
0.40
0.60
0.80
1
Normalized theta

Figure 4.12 Human Evolutionary Genetics, 2nd ed. (© Garland Science 2014)

JKU

Discovering and assaying variation at microsatellites

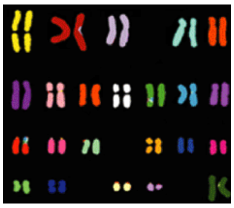
- PCR based assays:
  - □ using flanking primers in proximity of the repeat and then distinguish different PCR product sizes
  - □ Detection of different product sizes: via capillary electrophoresis (fluorescently labelled primer)
  - □ Multiplexing: simultaneous separation and detection of more than one microsatellite region (each primer has its own fluorophore



JKU

Assessing structural variation - microscopy

- Chromosome painting
  - □ Chromosomal rearrangements are detected via cytogenic analysis (chromosome painting). Resolution of 2-3Mb
  - □ information about the per chromosome copy number of a variant
  - □ are CNVs on one chromosome or different chromosomes?
- Fiber FISH
  - □ Fiber FISH demonstrates copy number of sequences on individual chromosomes.
  - □ High-resolution fiber FISH validation of copy number estimates for the gene AMY1



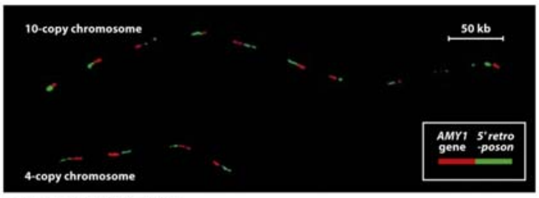
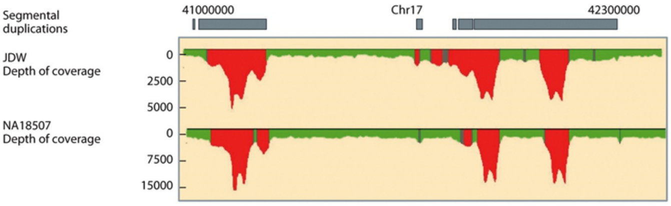


Figure 4.19 Human Evolutionary Genetics, 2nd ed. (© Garland Science 2014)

JKU

Assessing structural variation- sequencing

- Sequencing strategies to identify most CNV (differences in read-depth)
  - □ unusual read-depth
  - □ discrepant separation between read-pairs
  - □ reads that span a breakpoint and thus map to two distinct locations in the reference sequence.



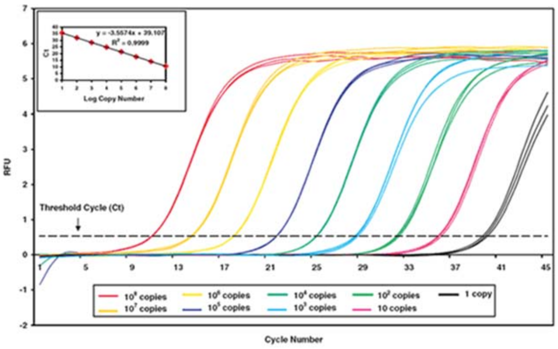
Red indicates regions of excess read-depth (mean + 3 standard deviations; s.d.); gray, regions of intermediate read-depth (mean + 2 s.d. and – 3 s.d.); green, regions of normal read-depth (mean ± 2 s.d.).

JKU

Assessing structural variation-with PCR

- Quantitative PCR (qPCR)
  - □ genomewide methods indicate the presence of a CNV, but not the precise number of copies of a given CNV
  - □ one and two copies are easy to distinguisg compared to eg. four and five copies
  - □ qPCR determines the amount of input copies by using standards;
  - □ a region outside the CNV is simultaneously amplified as a normalization control

Fig. 2. Real-time PCR Amplification using HotStart-IT™ Probe qPCR Master Mix with UDG (PN 75764).



Outline for today-part 1

- Methods to detect genomic variation
- Sequencing Projects
- Understanding a genome sequence
- Structural features

Animal sequencing projects

Fruit fly
Drosophila melanogaster

Rat
Rattus norvegicus

Mouse *Mus*
musculus

Zebrafish *Danio*
rerio

Domestic Gallus
gallus

Chimpanzee
Pan troglodytes

Rhesus macaque
Macaca mulatta

Domestic dog
Canis familiaris

Tammar wallaby
Macropus eugenii

Pufferfish
Tetraodon nigroviridis

Purple sea urchin
Strongylocentrotus purpuratus

Sea squirt
Ciona savignyi

Platypus
Ornithorhynchus anatinus

Mosquito
Anopheles gambiae

Honeybee
Apis mellifera

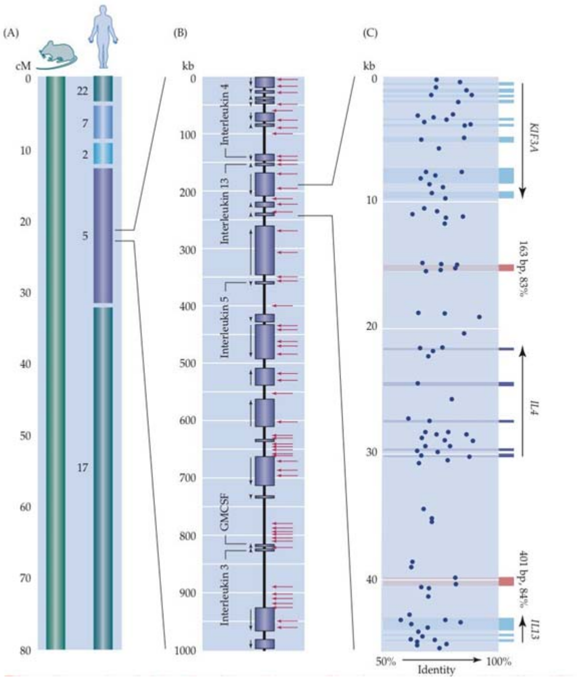
Non-human Primate Genome Sequencing

- Primates that have been sequenced: chimpanzee, gorilla, gibbon, orangutan

- More primitive primates: galago and mouse lemur. Model organism—Rhesus macaque

Functional genomics

Alignment prepared using PipMaker
http://pipmaker.bx.psu.edu/pipmaker



Rodent genome projects

- Sequencing model organisms
  - Large number of mutant strains
  - Potential of whole genome mutagenesis
  - 100 strains with well characterized genealogy—pedigrees for association studies
  - Functional genomics—comparison of conservation of proteins

Other vertebrate models

- ■ Pigs:
  - □ Specific genes unique in pigs:
    - ● olfactory receptor genes, which explains the pig’s sensitive sense of smell and ability to seek out things like truffles.
    - ● Taste receptors, however, have had additions over time making them weaker, which is one reason pigs can tolerate food that humans find unappealing.
  - □ The pig genomes also give an understanding of how pigs can become even better models for human health. Already, pigs are used as models for human hearts and cardiovascular systems. Their physiology and organ size is more similar to humans than most animals.

Other vertebrate models

- ■ Dogs: Model for complex diseases
  - □ Large range of phenotypes in inbred lines
    - ● Dermatitis
    - ● Parasite infection
    - ● Behavioral disorders
    - ● Heart conditions
  - □ Eg. Igf1 (insulin-like growth factor) contributes to the small size in chihuahuas
- ■ Zebrafish: rapid and transparent embriogenesis
  - □ Fluorescent labels: development of heart, eye, nervous system

Invertebrate model organisms

- Worm (C. elegans)—sequenced 1998
  - □ www.wormbase.org
- Fly (D. melanogaster)—sequenced 2000
  - □ www.flybase.org
- 19,000 vs 13,500 genes in the fly compared to the worm
- There is no relationship between gene number and tissue complexity
- Invertebrate genome projects: identifiable mutations for every gene of the genome can be obtained
  - □ iRNA in the food
  - □ 85% of genes of worms have been knocked-out

Plant genome

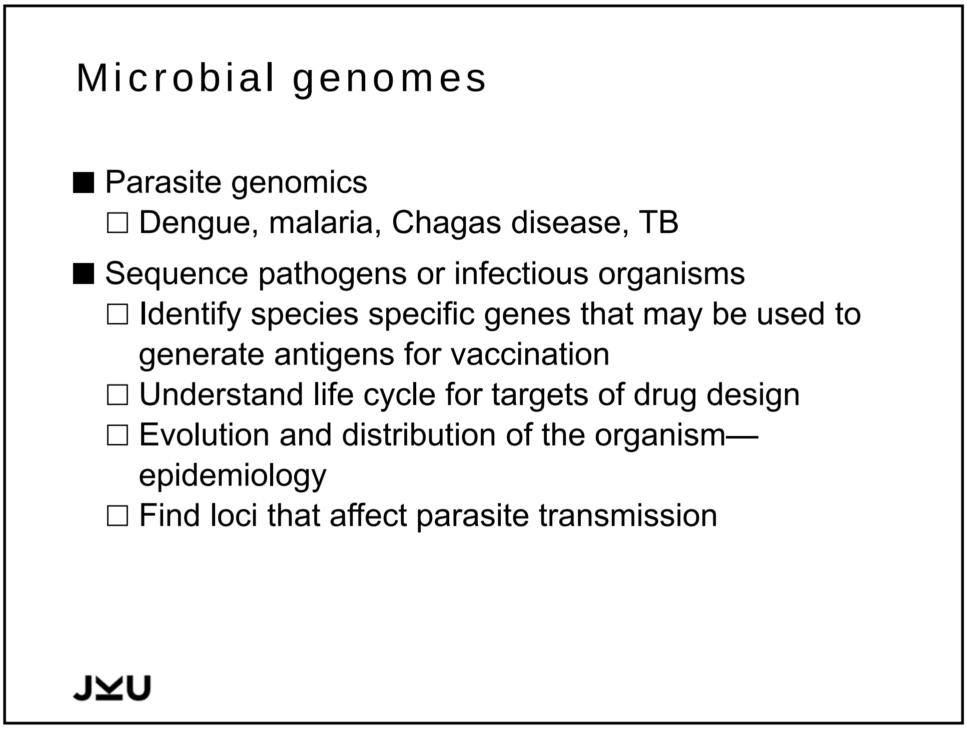
- Arabidopsis thaliana sequenced in 2000
  - □ Contains twice as many genes as the same euchromatic portion in the fly (25,000)
  - □ Two rounds of whole genome duplication followed by re-shuffling and gene loss: 2-3 gene duplicates—assigned to 11,000 gene families

Plant genome

- Most genes are of organelle descent and lost they targeting signal
- Plant-specific genes:
  - Cell wall biosynthesis
  - Transport proteins: nutrients, ions, metabolites moved between cells
  - Photosynthesis
  - Plant turgor, gravitrophic and phototrophic response genes
  - Secondary metabolites (cytochromes)
  - Pathogen resistant genes

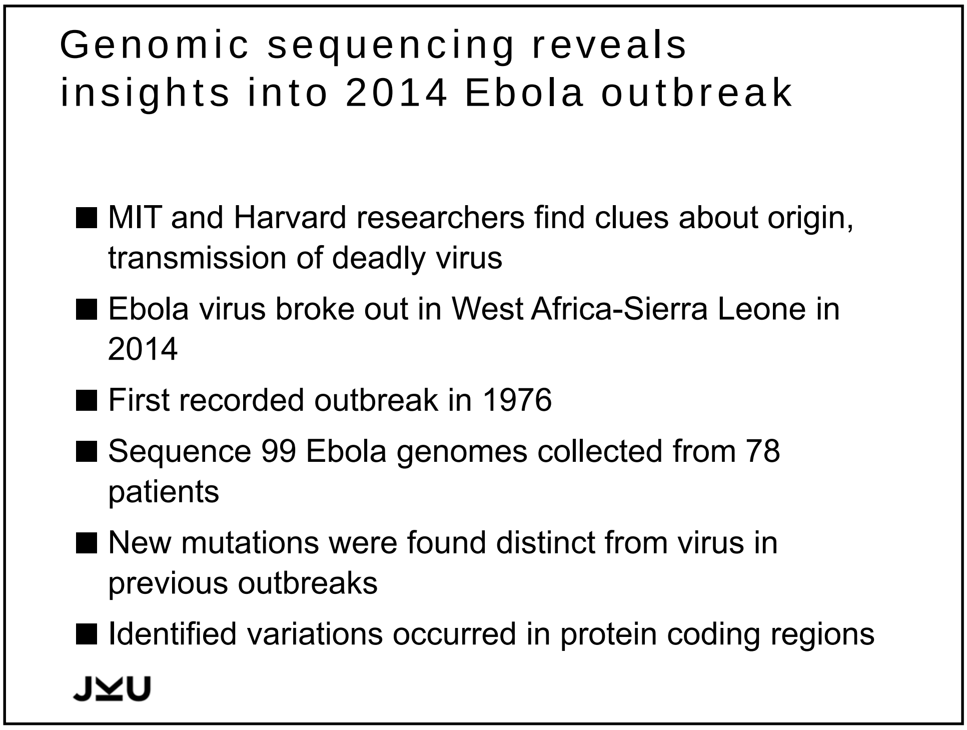
Microbial genomes

- S. cerevivisiae (yeast)—model organism
  - 6000 predicted genes
  - Gene duplications
  - Functional genomics (targeted mutagenesis, protein interaction with the yeast-two hybrid system)



Microbial genomes

- ■ Parasite genomics
  - □ Dengue, malaria, Chagas disease, TB
- ■ Sequence pathogens or infectious organisms
  - □ Identify species specific genes that may be used to generate antigens for vaccination
  - □ Understand life cycle for targets of drug design
  - □ Evolution and distribution of the organism—epidemiology
  - □ Find loci that affect parasite transmission

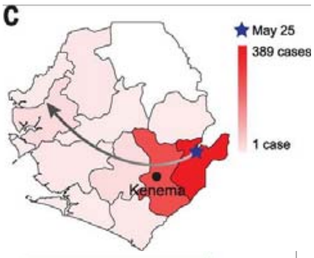


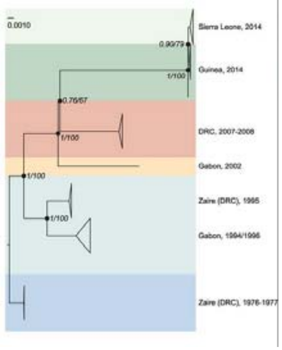
Genomic sequencing reveals insights into 2014 Ebola outbreak

- ■ MIT and Harvard researchers find clues about origin, transmission of deadly virus
- ■ Ebola virus broke out in West Africa-Sierra Leone in 2014
- ■ First recorded outbreak in 1976
- ■ Sequence 99 Ebola genomes collected from 78 patients
- ■ New mutations were found distinct from virus in previous outbreaks
- ■ Identified variations occurred in protein coding regions

Genomic sequencing reveals insights into 2014 Ebola outbreak

- Variations occurred in protein coding regions can be used the following way:
- trace the transmission path and evolutionary relationships of the samples
  - □ The lineage responsible for the outbreak diverged from the Middle African version of the virus within the last ten years and spread from Guinea to Sierra Leone by 12 people who had attended the same funeral.



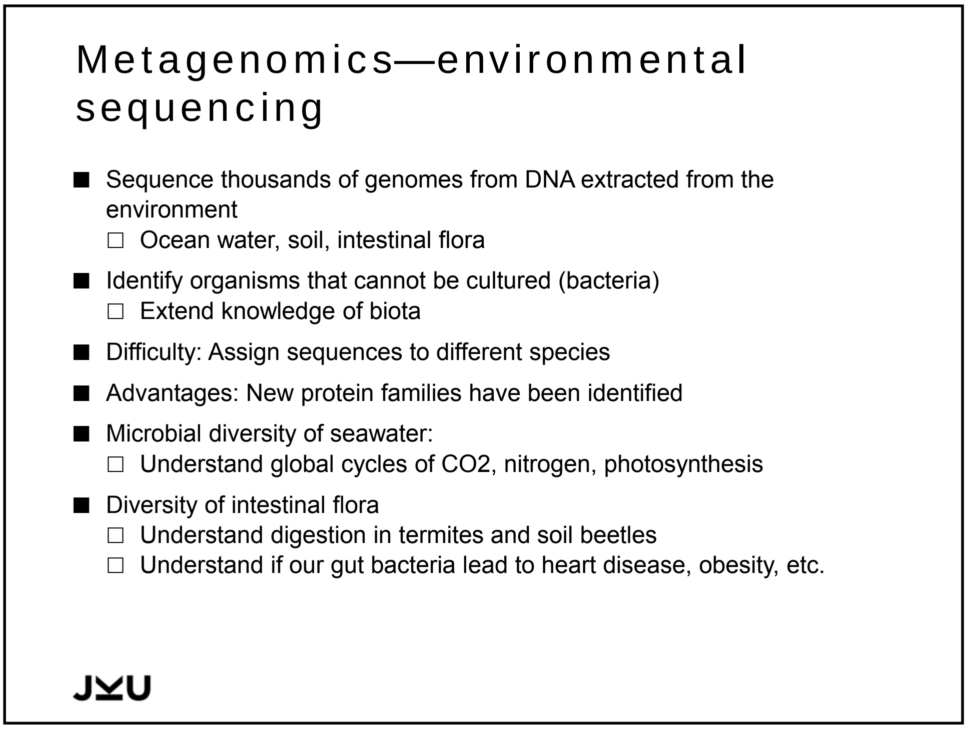


JKU

Genomic sequencing reveals insights into 2014 Ebola outbreak

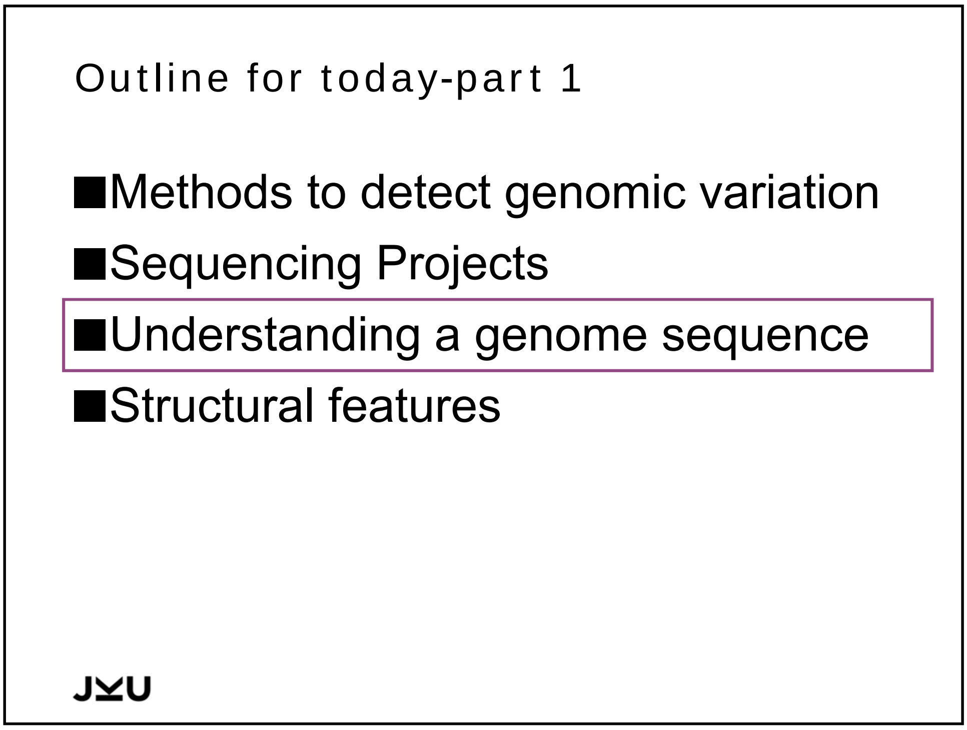
- Variation also affected diagnostics (primers sequences)
  - □ In May population was negative for PCR diagnostic test
  - □ Were these false negatives?
  - □ Design new primers with sequence of the whole genome (deposited in NCBI)
- Variation can be used for potential targets for future diagnostics, vaccines, and therapies
  - □ “Although we don’t know whether these differences are related to the severity of the current outbreak, by sharing these data with the research community, we hope to speed up our understanding of this epidemic and support global efforts to contain it.”
  - □ Read more at: Gire, SK, Goba, A et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science.
  - □ Tragically, five co-authors, who contributed greatly to public health and research efforts in Sierra Leone, contracted EVD and lost their battle with the disease before this manuscript could be published

JKU



Metagenomics—environmental sequencing

- ■ Sequence thousands of genomes from DNA extracted from the environment
  - □ Ocean water, soil, intestinal flora
- ■ Identify organisms that cannot be cultured (bacteria)
  - □ Extend knowledge of biota
- ■ Difficulty: Assign sequences to different species
- ■ Advantages: New protein families have been identified
- ■ Microbial diversity of seawater:
  - □ Understand global cycles of CO2, nitrogen, photosynthesis
- ■ Diversity of intestinal flora
  - □ Understand digestion in termites and soil beetles
  - □ Understand if our gut bacteria lead to heart disease, obesity, etc.



Outline for today-part 1

- ■ Methods to detect genomic variation
- ■ Sequencing Projects
- ■ Understanding a genome sequence
- ■ Structural features

Understanding a genome sequence

- Genome Annotation:
  - □ Genes
  - □ Regulatory sequences
  - □ Non-protein coding genes
  - □ Repetitive sequences
  - □ Segmental duplications
  - □ Non-genic sequences

The biological dogma: From genes to proteins

- Genes, transcription and translation

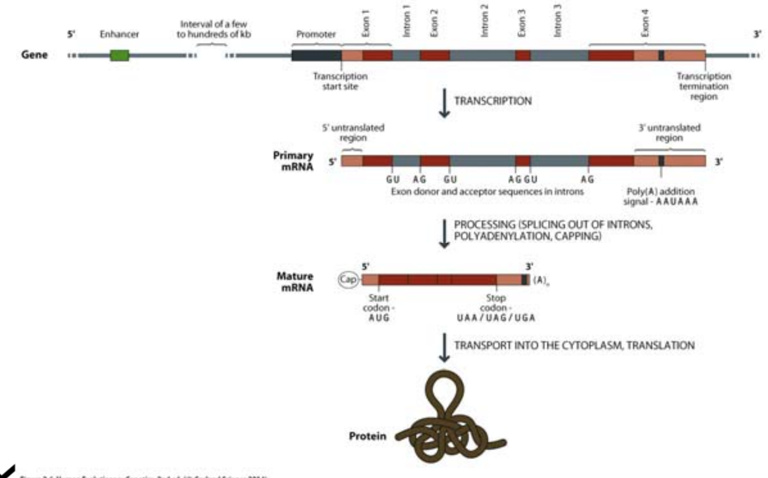


Figure 2.6 Human Evolutionary Genetics, 2nd ed. (© Garland Science 2014)

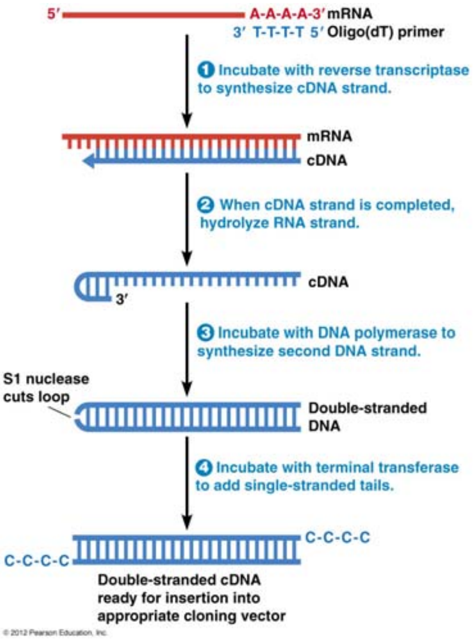
How are genes identified?

- EST sequencing
  - proof that a genomic fragment is transcribed
- Ab Initio Gene Discovery
  - Computational inferences on genomic sequences

From mRNA to DNA

Sequence mRNA

- Prepare mRNA library (represent mRNA expressed in the tissue)
- mRNA gets converted into cDNA
- cDNA is sequenced



1. Incubate with reverse transcriptase to synthesize cDNA strand.
2. When cDNA strand is completed, hydrolyze RNA strand.
3. Incubate with DNA polymerase to synthesize second DNA strand.
4. Incubate with terminal transferase to add single-stranded tails.

Double-stranded cDNA ready for insertion into appropriate cloning vector

© 2012 Pearson Education, Inc.

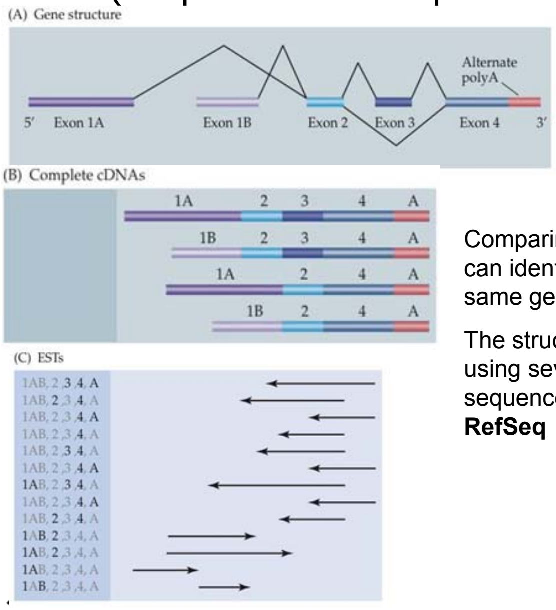
Finding genes

- EST sequencing
  - mRNA→cDNA→cDNA library→expressed sequence tag (partial sequence of a cDNA)
    - The frequency of the majority of transcripts is 1:10000
    - sequencing ~500clones=0.01%-1% of total transcripts
    - cDNA library needs to be normalized to account for expression differences in time and space

cDNA libraries

- Human Gene expression:
  - Atlases of human gene expression
    - Cancers
    - Splice variants
    - Tissues
    - Cell lines
- Clone sequences can be downloaded from NCBI (dbEST) or from Ensembl
- Prior to the human genome project the number of EST exceeded 100,000 (4x the number of genes)
- Why are there so many ESTs?
  - Average human gene: 10-15 exons; 3 different proteins from alternative splicing

EST (Expressed sequences tag)



Comparing EST with genomic DNA can identify EST representing the same gene

The structure of a gene is defined using several complete cDNA sequences of high quality libraries—**RefSeq**

Ab Initio Gene Discovery

- Algorithm recognizing features common in protein-coding regions
  - Identify Open Reading Frames (ORFs)
    - Codon bias similar to the species studied
    - Transcription/translational initiation motifs
    - 3’ polyadenylation sites
    - Splicing consensus sequences between introns and exons

- Higher eukaryotes with long introns and extensive intergenic regions (or bizarre genes within the introns of other genes) need special tweaks for *ab initio* discovery
- Genie, Genscan HMMgene are programs to find ORFs

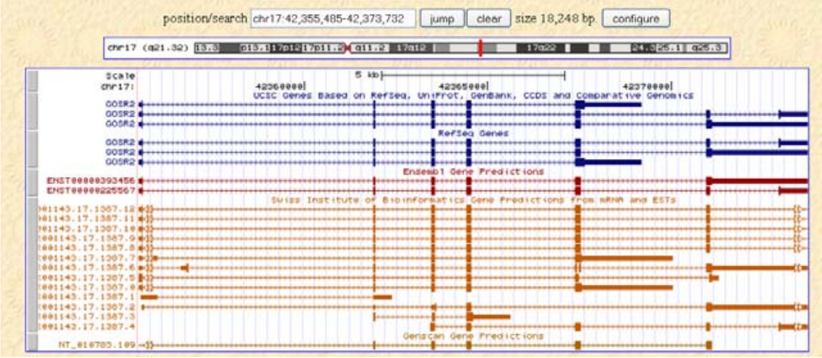
Ab Initio Gene Discovery

- Direct evidence must be provided to confirm a predicted gene
  - □ Reference to an annotated cDNA
  - □ Match one or more EST from the same organism
  - □ Conservation (Similarity to other proteins or nucleotide sequences across organism)
  - □ Domain predictions matches PFAM database
  - □ Promotor sequences, transcription binding sites, CpG island

JKU

How reliable is Ab Initio Gene Discovery?

- Correct identification of only 75% of exon-intron boundaries
- Identification of 80-90% of all true genes (<10% false positive rate)
- Some genes were identified after the introduction of a mutation causing a severe phenotype

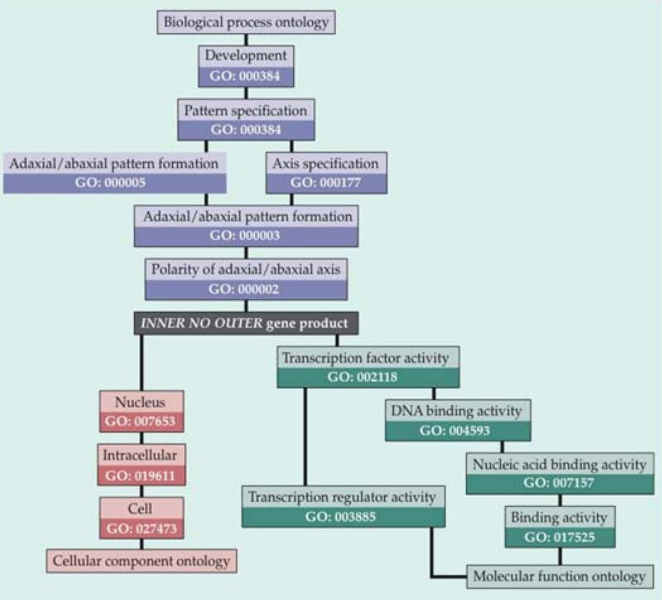


position/search chr17:42,355,485-42,373,732 size 18,248 bp.

JKU

Gene Ontology consortium

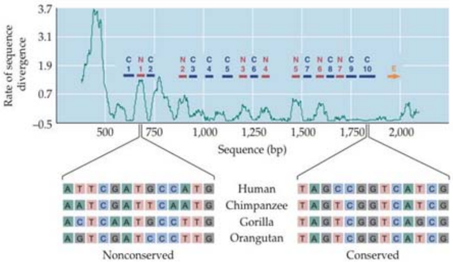
- Gene annotation based on standard vocabulary to classify proteins based on function
  - □ Cell function (process that protein is involved)-**What?**
    - ● Cell growth, division, replication
  - □ Molecular function (biochemical pathway)-**How?**
    - ● Enzyme, nucleic binding, transcription, etc.
  - □ Cellular component the protein is active-**Where?**
    - ● Cell surface, Golgi, nucleus


Biological process ontology
Development
GO: 000384
Pattern specification
GO: 000384
Adaxial/abaxial pattern formation
GO: 000005
Axis specification
GO: 000177
Adaxial/abaxial pattern formation
GO: 000003
Polarity of adaxial/abaxial axis
GO: 000002
INNER NO OUTER gene product
Transcription factor activity
GO: 002118
Nucleus
GO: 007653
DNA binding activity
GO: 004593
Intracellular
GO: 019611
Nucleic acid binding activity
GO: 007157
Cell
GO: 027473
Transcription regulator activity
GO: 003885
Binding activity
GO: 017525
Cellular component ontology
Molecular function ontology

GO network for a typical plant gene.
(After figure found on http://www.geneontology.org/go.nodes.html.)

Regulatory sequences

- Conserved non-coding sequence
- 5’ to the transcription site, intronic or located at the 3’ end
- Phylogentic footprinting: compare sequences of different species and test for conservation
- Phylogenetic shadowing: apply model of sequence evolution to sequences of related species



JKU

Regulatory sequences

- Identification of candidate regulatory regions can be confirmed by binding nuclear extracts (gel shift assays)
- Identifying short regulatory motifs
  - Look for motifs that are repeated in the genome
  - Look for candidate proteins that might recognize that motif
  - Test by Immuno-precipitation if motif is bound by predicted protein

JKU

Non-Coding Genes (do not result in proteins)

- ■ Identification of functional RNAs
- ■ Challenges
  - ☐ No poly adenylation signal (no standard cDNA library)
  - ☐ No codon divergence but secondary structure (no sequence similarity to rely on)
  - ☐ Very little information on the function and distribution of non-coding RNA (ncRNA)

JKU

Non-coding RNA

- ■ tRNA (transfer RNA)
  - ☐ Involved in protein synthesis
  - ☐ ~500 classes
  - ☐ Dispersed in large clusters
  - ☐ tRNA content in the human genome

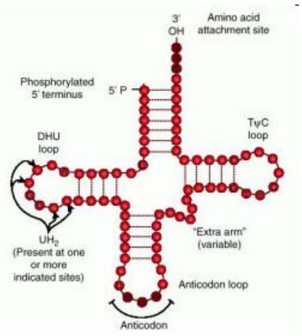
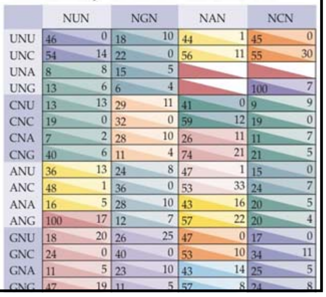
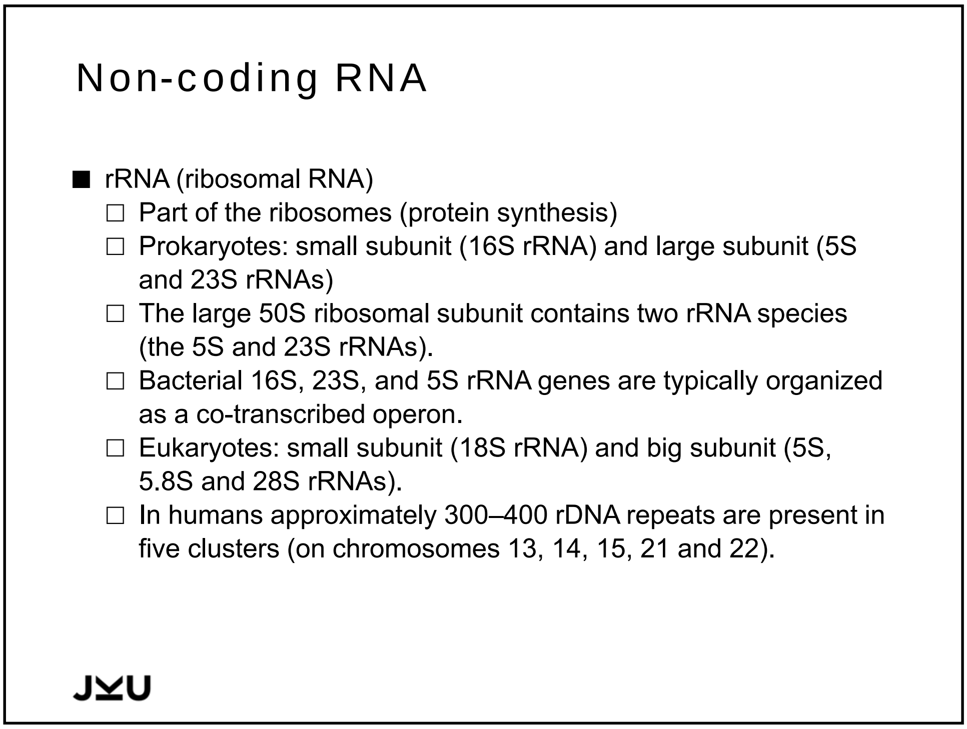

3' OH
Amino acid attachment site
Phosphorylated 5' terminus
5' P
TψC loop
DHU loop
"Extra arm" (variable)
UH_2 (Present at one or more indicated sites)
Anticodon loop
Anticodon

Figure 2.16 Transfer RNA content in the human genome. (A) The table indicates the percentage of codons for each color-coded amino acid below the diagonal, and the number of tRNAs with the associated anticodon above the diagonal in each box. (B) The plot shows that codon bias (expressed as the percentage of the particular amino acid that is encoded by each codon) is only partially correlated with the number of tRNAs with the complementary anticodon. Thirteen possible codons are not associated with exactly complementary tRNAs, and so must be decoded by "wobble" anticodons.



	NUN	NGN	NAN	NCN
UNU	46 0	18 10	44 1	45 0
UNC	54 14	22 0	56 11	55 30
UNA	8 8	15 5		
UNG	13 6	6 4		100 7
CNU	13 13	29 11	41 0	9 9
CNC	19 0	32 0	59 12	19 0
CNA	7 2	28 10	26 11	11 7
CNG	40 6	11 4	74 21	21 5
ANU	36 13	24 8	47 1	15 0
ANC	48 1	36 0	53 33	24 7
ANA	16 5	28 10	43 16	20 5
ANG	100 17	12 7	57 22	20 4
GNU	18 20	26 25	47 0	17 0
GNC	24 0	40 0	53 10	34 11
GNA	11 5	23 10	43 14	25 5
GNG	47 19	11 5	57 8	24 8

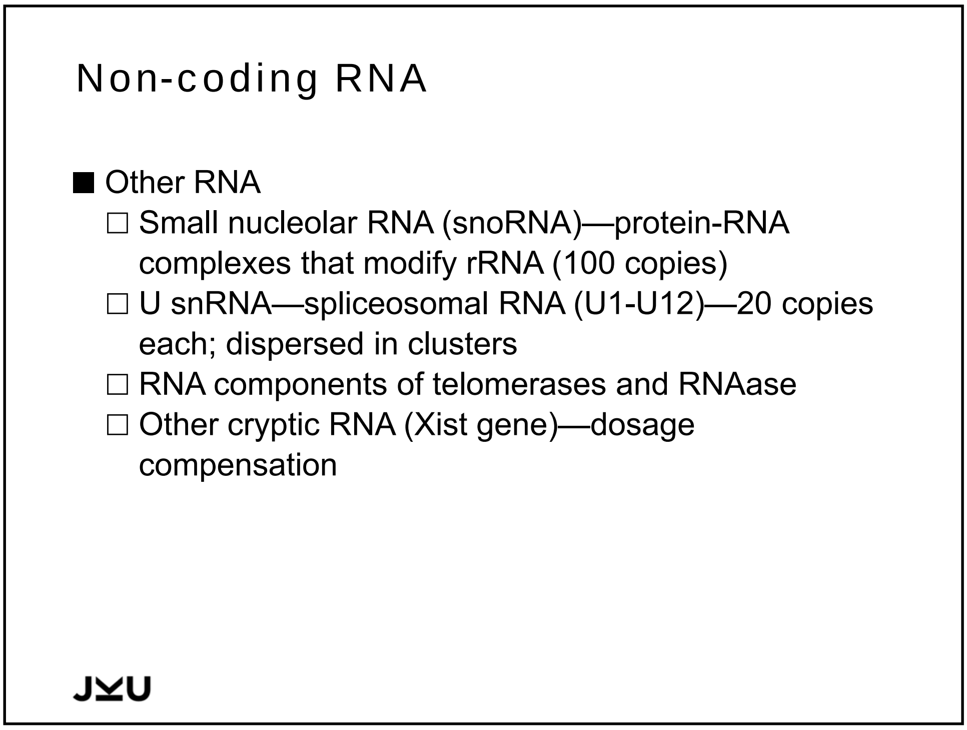
JKU



Non-coding RNA

- rRNA (ribosomal RNA)
  - □ Part of the ribosomes (protein synthesis)
  - □ Prokaryotes: small subunit (16S rRNA) and large subunit (5S and 23S rRNAs)
  - □ The large 50S ribosomal subunit contains two rRNA species (the 5S and 23S rRNAs).
  - □ Bacterial 16S, 23S, and 5S rRNA genes are typically organized as a co-transcribed operon.
  - □ Eukaryotes: small subunit (18S rRNA) and big subunit (5S, 5.8S and 28S rRNAs).
  - □ In humans approximately 300–400 rDNA repeats are present in five clusters (on chromosomes 13, 14, 15, 21 and 22).

JKU



Non-coding RNA

- Other RNA
  - □ Small nucleolar RNA (snoRNA)—protein-RNA complexes that modify rRNA (100 copies)
  - □ U snRNA—spliceosomal RNA (U1-U12)—20 copies each; dispersed in clusters
  - □ RNA components of telomerases and RNAase
  - □ Other cryptic RNA (Xist gene)—dosage compensation

JKU

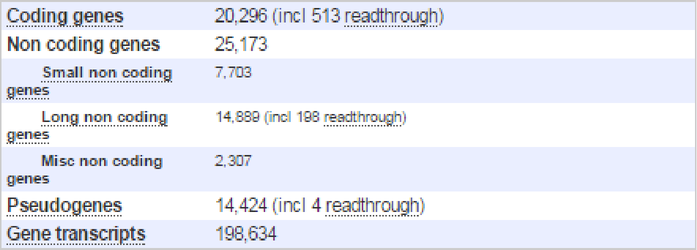
microRNA or interference RNA (iRNA)

- Short hairpin RNA (miRNA)
- 21-23 nucleotides in length
- miRNA start as mRNA but are not translated into proteins, but processed into a stem-loop and transformed to a single stranded RNA
- Partial or fully complementary to one or more mRNA
- Bind to 3'untranslated regions of mRNAs and function in gene regulation either by repressing translation or promoting mRNA degradation
- >1000 known miRNAs from numerous organisms

JKU

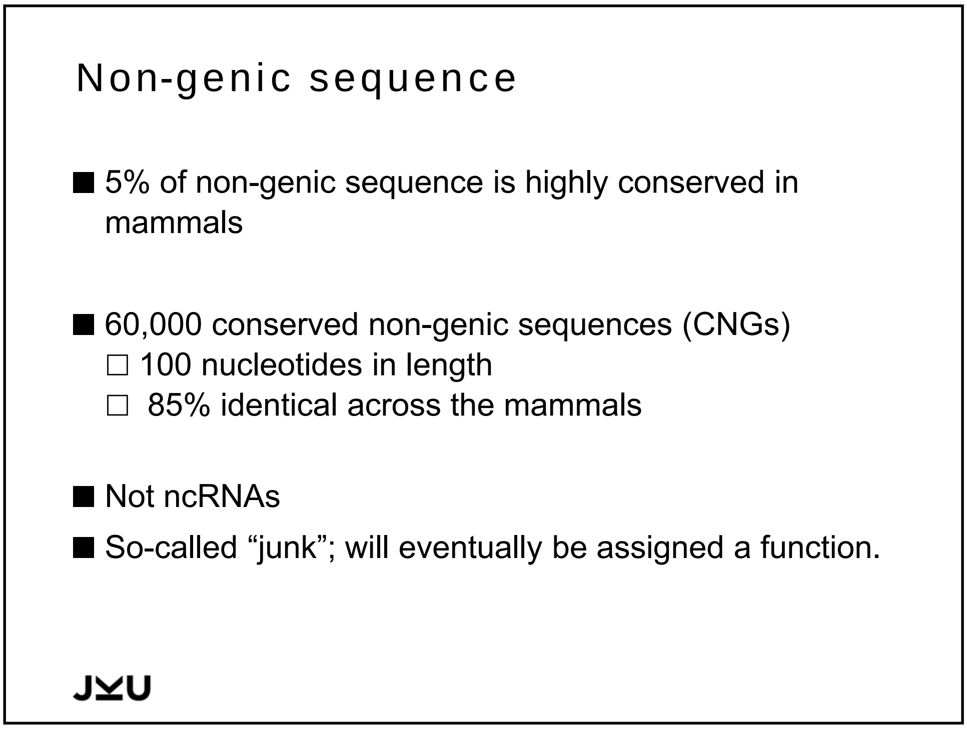
Number of non-coding genes

http://www.ensembl.org/Homo_sapiens/Info/Annotation
Genebuild: Jul2015



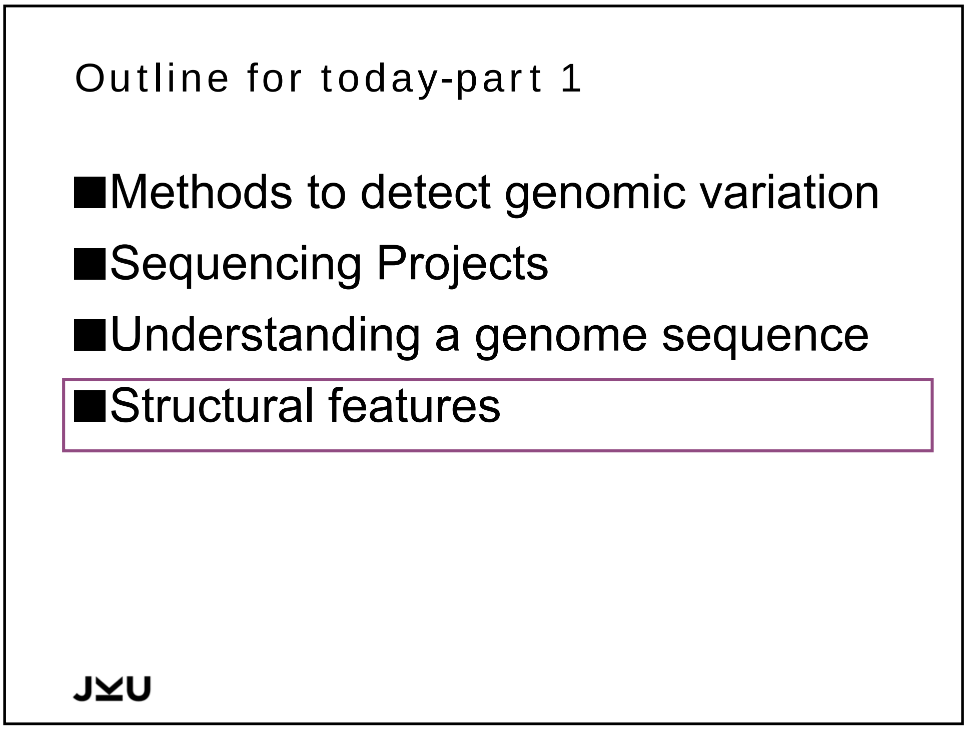
Coding genes	20,296 (incl 513 readthrough)
Non coding genes	25,173
Small non coding genes	7,703
Long non coding genes	14,889 (incl 198 readthrough)
Misc non coding genes	2,307
Pseudogenes	14,424 (incl 4 readthrough)
Gene transcripts	198,634

JKU



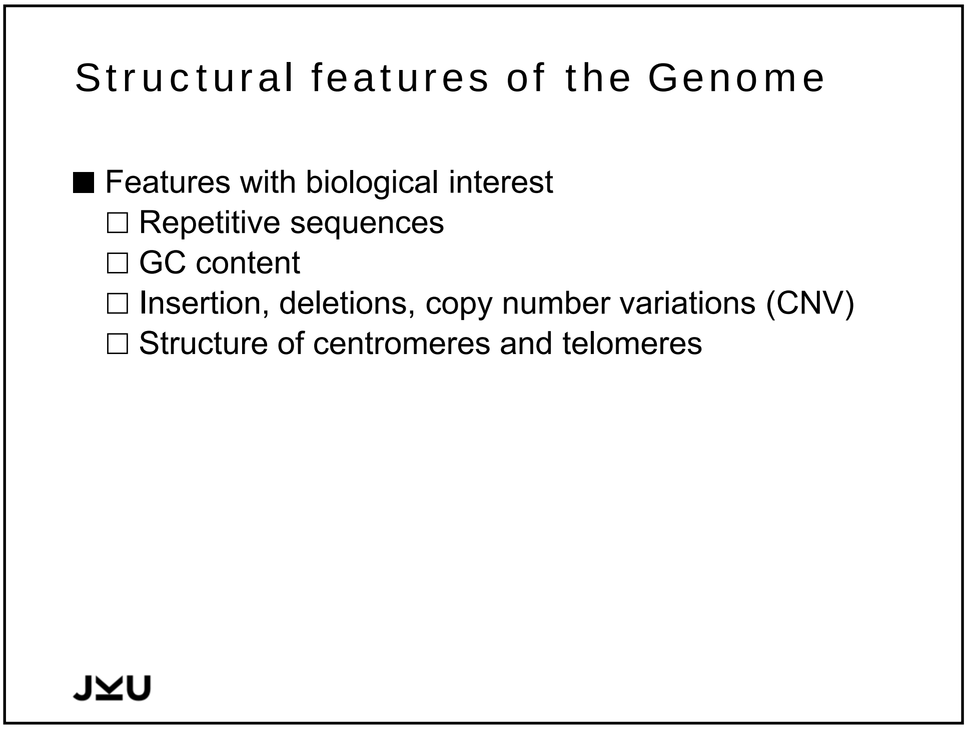
Non-genic sequence

- 5% of non-genic sequence is highly conserved in mammals
- 60,000 conserved non-genic sequences (CNGs)
  - □ 100 nucleotides in length
  - □ 85% identical across the mammals
- Not ncRNAs
- So-called “junk”; will eventually be assigned a function.



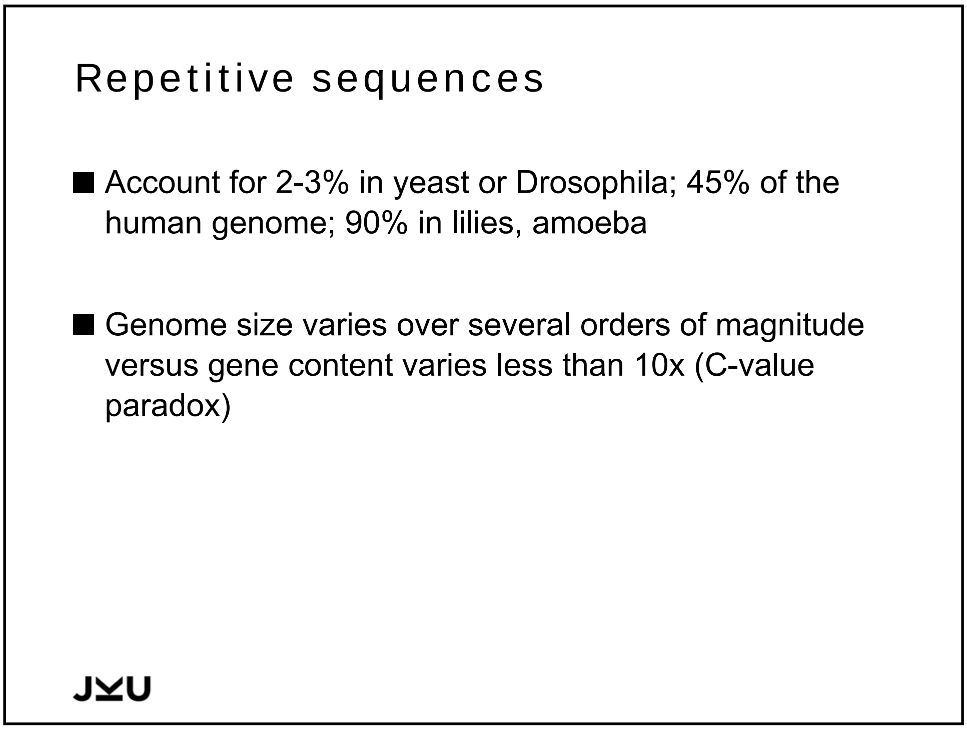
Outline for today-part 1

- Methods to detect genomic variation
- Sequencing Projects
- Understanding a genome sequence
- Structural features



Structural features of the Genome

- Features with biological interest
  - □ Repetitive sequences
  - □ GC content
  - □ Insertion, deletions, copy number variations (CNV)
  - □ Structure of centromeres and telomeres



Repetitive sequences

- Account for 2-3% in yeast or Drosophila; 45% of the human genome; 90% in lilies, amoeba
- Genome size varies over several orders of magnitude versus gene content varies less than 10x (C-value paradox)

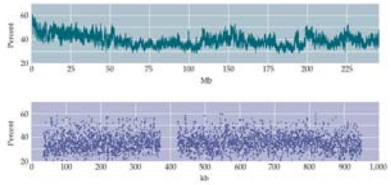
GC content

- ■ Genomes have wide variation in GC content
  - □ Humans: 40-60%; Plasmodium 10-30%
  - □ GC content depends on:
    - ● temperature of environment
    - ● Levels of methylation
    - ● Transposon activity

JKU

GC content

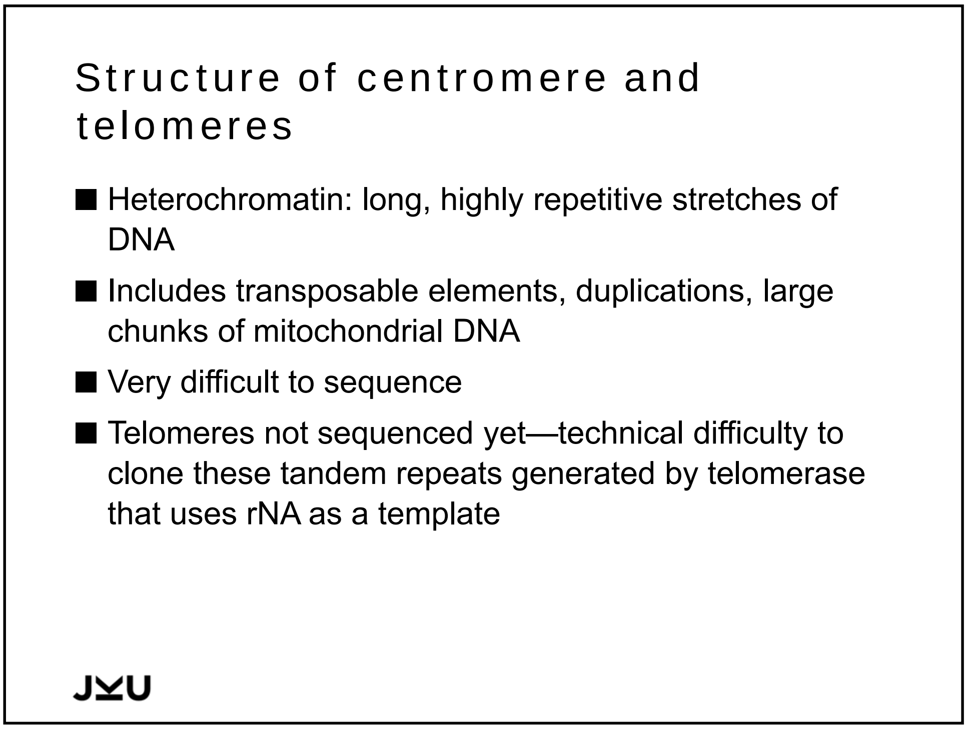
- □ GC isochores: regions with different GC content (~30% difference—karyotypic bands)



Distribution of GC content along human chromosome 1.

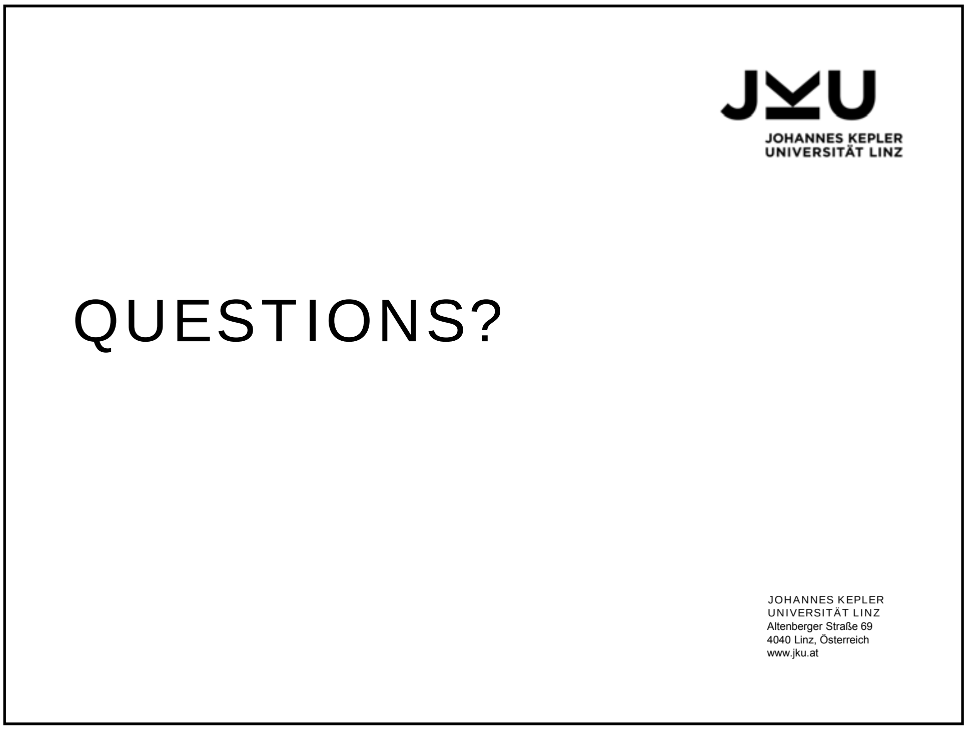
- □ CpG sites underrepresented in mammalian genomes, except in CpG islands (0.5-2kb) found upstream of genes

JKU



Structure of centromere and telomeres

- Heterochromatin: long, highly repetitive stretches of DNA
- Includes transposable elements, duplications, large chunks of mitochondrial DNA
- Very difficult to sequence
- Telomeres not sequenced yet—technical difficulty to clone these tandem repeats generated by telomerase that uses rNA as a template



JOHANNES KEPLER
UNIVERSITÄT LINZ

QUESTIONS?

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