21.5.2012

Describe briefly THREE of the following topics:

1. **Comparative genomics and its applications**

   The use of information obtained from the study of oen genome to make inferences about the positions and functions of genes in another genome. Comparative genomics starts by making alignment of two or more DNA sequences and proceeds by observing if there are orthologous sequences in the aligned DNA and if so, what is the level of their conservation.

Common features of two sepqieces are encoded by DNA that is conserved between species. Likeqise, the sequences controlling the genes that are regulated similarly are also conserved, as also are other types of functional sequences not encoding proteins.

Is a sequence conserved beyond neutral expectations? If yes -> under negative selection, indicating that the sequence is functional. Two organisms with relatively recent common ancestor will have genomes that display species-secific on the common ancestral genome. Comparison of genomes at different phylogenetic distances are appropriate to address different questions.

You find ultra conserved regions in the genomes. Other than the protein coding parts, human genome contains several sequence elements that are conserved. They are obsolutely conserved and can be found in rats and mice. They are often close to genes that encode transcriptional regulation.

2. Tandem repeats in the human genome

- Most of the human genome is extragenic or non-coding, only 3% is protein coding. The types of repetitive sequences: satellites, interspersed, retrotransposons. Tandem repeats are a pattern of nucleotides that repeats directly adjacent to each other.

Tandem repeats: Satellite DNA is located on heterochromatic regions of chromosomes, especially on centromeres. Repeats if variable, but always short (5-200), but the amount of repeptitions is high and the actual size of satellistes is >100kb. In humans satellites have a detectable repetitive sequence. Satellite might have functions like take part in heat shcok response. Humans have three satellites that can be separated by density gradient centrifugation, and alpha and beta satellites that go with bulk DNA.

- alpha-satellites centromeric in all choromosomes

- beta-satellites centrometic in some chromosomes

- Satellite 1 centromeric in most

- Sat 2/3 most chromosomes

Minisatellites: Very short reperititve sequences that is the same to all humans. Short sequences in telomeres/telomeric regions and VNTR regions that are regions of DNA where different humans have different amounts of short repeats.

Microsatellites: Randomly in the genome, no CG-repeats. mono- or Dinucleotide repeats that repeat 10-20 times. Mostly generated by replication slippage.

3. Reverse genetics in analysis of gene function

Reverse genetics begins from genome sequencing information, and aims to understand the function of the gene and protein.

- Transgenic Organisms are essential to reverse genetics.

- Answers the questions "What phenotype are as a result of particular gene."

- To identify the influence a sequence has on the phenotype, researchers engineer a change or disruption in the DNA. After researcher can look for the effect of such alterations in the whole organism.

    - Directed deletions/point mutations

    - Gene silencing (RNAi)

    - Interference using transgens (Knockout, overexpression studies)

4. Microarrays; principle and use in defining regulons

Can detect expression of 1000s of genes simultaneously. Can be used for indentification of complex genetic diseases or toxicology studies, pathogen analysis, or differing expression levels overtime (disease state) or between tissues. Answers the questions: In which tissues your genes of interest are expressed? When are they switched on? How do they respond to disease, hormone, stress…?

Small solid disc with wells where thousands of sample genes are immobilized. Can be for oligonucleotides (25mers) or from cDNA libraries.

Fluorescent probes are prepared from two mRNA sources to be compared. Cy3 green is one and Cy5 red is the other. Probes are mixed and washed over the microarray. Ech probe is excited using a laser and its fluorescence at each element detected with scanning microscope. Green is control DNA where either DNa or cDNA derived from normal tissue is hybridized to target DNA. Red is sample DNA where tiehr DNA or cDNA is derived from diseased tissue and hybridized to target DNA. Yellow is when both are present. Black is where neither is present.

1. Post-translational modifications of proteins and how to analyze them

Many modifications are known, their biological function is to allow a stable 3D structure, allow solubility/insolubility, regulation, enabling protein-prtein interaction, localization. Some of these modifications: Phosphorylation, Methylation, Acetylation, Ubiquination…

In the past with Edman Degradation and now with Mass Spectrometry.

Goal: Determination of exact molecular mass of a protein followed by comparison with mass predicted from aa-sequence. Chemical or enzymatic removal of modifications monitored by MS. Sequencing of the modified peptide by MS/MS.

2. **What affects the genome size?**
    - Amount of DNA is not releated to the complexity of the species, but minimum genome size is related to complexity of the species. Average mammal has a larger genome than all prokaryotes and most amphibians. Prokaryotic genomes can be smaller, because they have the option of changing plastids on top of their mandatory plastids and linear chromosomes. Eukaryotic genomes are variable in size and density. Eukaryotic genomes have a lot of repetitive sequences. Only 3% of the human genome is protein coding. Most of the genome is extragenic or non-coding. Size of introns increases when complexity increases. Eukaryotic genomes allow for alternative splicing.
    - Small scale changes on the genetic level, recombinatios, duplications.
3. **How to assess gene function**

Interspersed repeats in eukaryotic DNA

Interspersed repetitive DNA is generated through transposition. Genome contains different transposable elelements in numerous copies.

DNA transposons: Copy paste migration, mostly inactive (transposon fossil). About 3% of the genome. Mer1 and Mer2.

But mostly interspersed repeats are Retrotransposons. They are made by transcription and reverse transcription of existing retrotransposon. The generated cDNA is inserted back into the genome. Retrotransposons are related to retroviruses. There are long terminal repeat(LTR) and non-LTR retrotransposons.

- There are endogenous retroviruses (HERV) that contain gag and pol genes (cis-regulatory element in retrovireuses that facilitaties the mechanism of translation). Most are defective and haven't transposed in millions of years.
- Non-autonomous retroviral elements lack at least the pol gene.

- LINEs (no LTR) ~20% of the genome. Humans have three families of LINEs. They are actively transposing.

- SINEs. Do not encode for transposition but can transpose by borrowing an enzyme that another retroelement has produced. Dispersed over the chromosome. Most common SINe in the human genome is the Alu-element.

2. **Explain 2 large-scale methods for anlyzing gene expression at the transcriptional level**

Expression profiling experiments often involve measuring the raltive amount of mRNA expressed in two or more experimental conditions.

cDNA libraries. EST (expressed sequence tag). is a short sub-sequence of cDNA. They may be used to identify gene transcripts. an EST results from one-shot sequencing of a cloned cDNA. the cDNA used for EST generation are typically individual clones from a cDNA library. The resulting sequence is a short fragment of completmentary to mRNA. EST can be mapped to specific chromosome locations using physical mapping thechicues like FISH. Or if the genome has been sequences, you can align the EST sequence to that genome. (map).

Microarrays are being increasingly replaced by sequencing based moethods (NGS). Short tags assign the site of oroigin of the read, Identification and quantification of transcripts wihout prior knowledge of a particular gene. Information regarding alternative splicing.

Sage serial analysis of gene expression.

Prduces a snapshot of the mRNA population in the sample of interest. Sage provides quantative and comprehensive expression profiling in a cell population. Allows rapid, detailed analysis of thousands of samples.It can be used to detect novel transcripts with no prior knowledge about sequence.

It works by creation of short sequence tags that contain enough information to uniquely identify a transcript. Sequence tags are linked together to form long serial molecules that can be cloned and sequenced. Quantification of the numer of times a particular tag is observed provides the expression level of the corresponding transcript.

SAGE software searched genbank for matches to each tag. Can be used for transcriptome analysis.

SAGE vs Microarray

Sage: Detects 3’. Collects seq information and copy number. Sequncing error/quantative bias. Detects unknown transcripts.

Array: Targets various regions of the transcript. Fluorescent. Labelib bias and noise. Highly specific. Cheaper than sage.

3. **Difference between forward and reverse genetic screens**
    - Forward genetics is identifying the genotype that is responsible for a phenotype. Generating random mutations in an organism subsequent breeding, looking how often traits are inherited together.

4. **Proteomics as part of systems biology**

Methods: Mass Spectrometry (identification, post-translational modifications, PPI).

Mass spectrometry to identify proteins: protein mass, primary sequence, post-transaltional modifications. Database search using MS or MS/MS data. Peptide fingerprint.
Extract -> enrich -> quantify (2DE) -> digest (trypsin) -> mass spectrometry -> identify.

2. What is a transcription factor and how to study its function?
3. **RNA silencing**

    RNA Interference (RNAi) is post-transcriptional silencing, a biological process where RNA molecules inhibit gene experassion typically by causing the desctruction of specific mRNA molecules. There's microRNA and small interfering RNA. RNA are the direct product of genes and these small RNAs can bind to other specific mRNA, and either decrease or increase activity, or guiding for destruction. In eucaryotes the enzyme Dicer cleaves long dsRNA into shor dsRNA (siRNA) and it unwinds to passenger and guide strand. Guide is incorporated to RNA-induced silencing complex.

    RNA can be used to targeted silencing of specific genes wth synthetic RNA. It can be used for large-scale screens that systematically shut down each gene in the cell to identify components of some mechanism.

4. **For what applications can sequencing be used?**

    Sequence tells us what the gene/cell could possibly do.

Describe briefly the generation of classical and processed pseudogenes

Pseudogenes are dysfunctional relatives on genes that have lost their protein coding abilities or are otherwise no longer expressed in the cell. Pseudogenes often result from accumulation of multiple

mutations in a gene that is not mandatory for survival. Most have gene like features like promoters and splice sites, but are nonfunctional (f.ex. premature stop codon or frameshift).

Pseudogenes are characterizd by a combination of homology to a nknown gene and non-functionality.

Processed pseudogenes are retrotransposed. Sometimes retrotransposition can happen with a gene, not a retrotransposon. Portion of the mRNA of the gene transcript is reverse transcribed and incorporated back into chromosomal DNA. These pseudogenes lack promoters and introns, and have a poly-A tail.

Non-processed pseudogenes come from gene duplication. A copy of a functional gene may arise as a result of a gene duplication event and aquire mutations that cause it to become non functional. Duplicated pseudogenes have all the same characteristics of genes (exon/intron…). There can also be disabled genes. Genes that have accumulated mutations and can be deactivated or become non-functional.