GGB – Genetics, Genomics, Bioinformatics

(MAKE A LIST OF MODEL ORGANISMS WITH THEIR PROS AND CONS FOR GENETIC STUDIES.)

21.2.2017

(On your own, formulate a sentence which puts genetics and evolution into relation and send it at ggb@biol.ethz.ch. Time: 1 week.
My sentence: The totality of the gene pool mirrors the current environmental and social structure of the world.)

(FOR CASE STUDIES, UNDERSTAND HOW AND WHY THE RELEVANT QUESTION WAS ASKED AND WHAT METHODS WERE USED TO ANSWER THE SPECIFIC QUESTION.)

I know classical, 2nd and 3rd generation sequencing method. I know a lot of technical features of technologies and measures. I know what can be done in principle.

One of the main questions we will be discussing is how does one's genotype and the environment correlate with the phenotype. (**Reminder**: 1bp in 1000 differs when two humans are compared.)
Genotype + Environment ➔$^{?}$ Phenotype
gene + (environment) ➔ Phenotype

In evolution, the phenotype is selected. Natural selection cares next to nothing about the genotype, of course, the phenotype is correlated with the genotype. The phenotype is the defining element for selecting ultimately.

Aflu: telomeres, methylation – what was it again?

Wentworth Thompson wrote a book on growth and form, where the tried to explain from a mathematical point of view. He failed (why again?).
The correct correlation which he missed was: genotype => development => phenotype. He missed the development step.

Genes code for the building blocks of an organism

It is: gene products ➔ networks (syntax of the biological language) ➔ operating system
One understands the gene products very well, but the syntax/network is the black box for us.

Understanding syntax:
Organism-minus-1-gene + protein-minus-1-organism leads to words.
double mutant analyses lead to gene interactions.
Further approaches: genome-wide association studies, proteomics and transcriptomics, dynamic modelling.

Def. digital phenotypes: Step between genome and phenotype. They help to translate mere information to its implementation/translation (dt. Umsetzung) to the phenotype finally. Dynamic modelling is being applied for this. (biology and computer science has to work together).

Alpha-Actinin-3: Why loss of gene is an evolutionary gain (ACTN3)

Commentary:
loss of function: The gene is rendered ineffective due to different processes, such as deletion, missense mutation, nonsense mutation (addition of a stop codon that leads to a premature expression of the protein that does not work).
In this case, there is an evolutionary advantage for not to carry the ACTN3

Allele frequency: over a population, how many subjects carry a specific allele (given in per cent).

Mice are used in the experiment because firstly, mice can reproduce a lot faster and the homozygous etc. trees can be observed a lot faster. Secondly, conducting this experiment in humans will lead to a big pool of potential errors (many factors could play a disturbing role in the experiment). In a human, one would need to recruit identical twins. Environmental, dietary and other cultural factors can be controlled a lot easier in mice than in humans.

How can one be sure that the findings/results in mice are also relevant in humans (**Reminder**: mice are 200 millions years apart from humans evolutionary)?
For once, certain genes and their functions are conserved throughout time. Secondly, mice have the same or very similar structures like human beings that function and act analogously in both organisms. Thirdly, Darwin's theory of evolution etc.

Calcium ions are important for muscle contraction. The occurrence and concentration of Ca2+ plays an important role in the mitochondrial biogenesis.

The reuptake of Ca2+ in the skeletal muscle fibers leads to heat, which is better for an organism to adapt in a cold climate. The reuptake is improved by 3- to 4-fold with ACTN3.

Impaired mitochondrial function has been associated with obesity and the metabolic syndrome. So, it is somewhat risky what natural selection is doing.

(HAVE A LOOK AT IT AGAIN, BECAUSE I'M NOT SURE WHAT I WAS WRITING.)

27.2.2017

Assumptions of the Hardy-Weinberg Equilibrium:
Organisms are diploid, only sexual reproduction occurs, generations are non-overlapping, mating is random (panmixia), population size is large, allele frequencies are equal in the sexes, there is no migration, mutation or selection.

If the subject of interest does not fulfil these assumptions, the genotype frequency versus allele frequency will often differ drastically from the HWE expectation, which is a parabola graph.

What is clustering? Things that are alike will be grouped together (assume a random distribution which can be sorted according to pre-chosen criteria such as k-means, Euclidian average etc.)

Types of clustering: agglomeration and partition.

Agglomeration: bottom-up method, many variations, often used in gene expression data, time intensive for large data sets (n**2)

Partition: top-down method, splits data sets in a given number of clusters, methods are k-means and self-organizing maps etc., time efficient for large data sets (n).

28.2.2017

Questions – Muddiest points (=: MP)

Genetic variability: Regarding normal RefSeq, how does one define a RefSeq for a protein, when some parts are simply not that important?
Answer. There is variability in all genomes, but it's only a certain part of a sequence. One simply chooses a sequence and defines it as the RefSeq. Other sequences are mutations etc. or simply defined as "other sequences of the RefSeq".

Low complexity regions: Why is it important to filter out low complexity regions? How is a low complexity region defined?
Answer. We remove them, because they generate alignments that are "uninteresting" and they make it more difficult to identify the "interesting" ones.

How to handle large data sets

Typically, one has to visualize them in some way and then they need to be manipulated or assorted that is clustering.
It is useful to choose several clustering methods for a biological problem in order to check whether they all lead to the same conclusion. If for some reason, only a specific clustering method will be used, it must be explained very well, why only one. The conclusion must be robust based on the data sets.
If one is not sure what metric to use, use both (or all). The best-case scenario is that the biological interpretation is the same.

Agglomerative clustering only works one way. What else? Multiple sequence alignment: Progressive alignment sequence for example.

1st generation: Sanger sequencing method:

Ingredients: DNA template strand, polymerase, primers, triphosphate deoxynucleotides dATP, dGTP, dTTP, dCTP and dideoxynucleotides ddATP, ddGTP, ddTTP, ddCTP. The difference lies in the hydroxyl groups, where the 3' hydroxyl group is replaced by an H-atom in ddATP etc. As long as dNTPs are added, the elongation will not terminate. The addition of an ddNTP will terminate the elongation and one is left with a DNA product strand that is most of the time shorter than its template strand. With this method, one can identify the position of a certain base in the DNA, repeat this process, and puzzle it together to identify the complete base sequence of the DNA.

2nd generation: sequencing by synthesis (SBS):

Same ingredients as in sanger method, but there are fluorescently labelled chain terminating nucleotides and no dNTPs or ddNTPs.

Procedure: 1) The flow cell is flooded with a solution containing the four different fluorescently labelled nucleotides.
2) The polymerase uses the template strand as a guide to extend the primer with the complementary nucleotide. The terminating group on the newly incorporated nucleotide prevents the incorporation of additional nucleotides.
3) Any non-incorporated nucleotides are washed away.
4) the fluorescence signal of each template cluster of the sample is measured by fluorescence

microscopy and reveals which of the four nucleotides has just been incorporated.
5) Both the fluorophore and the terminating group are removed by chemical cleavage and washed away. This prepares the reaction for the next cycle.

SBS makes it possible to perform billions of parallel sequencing reactions in a flow cell the size of a microscope slide.

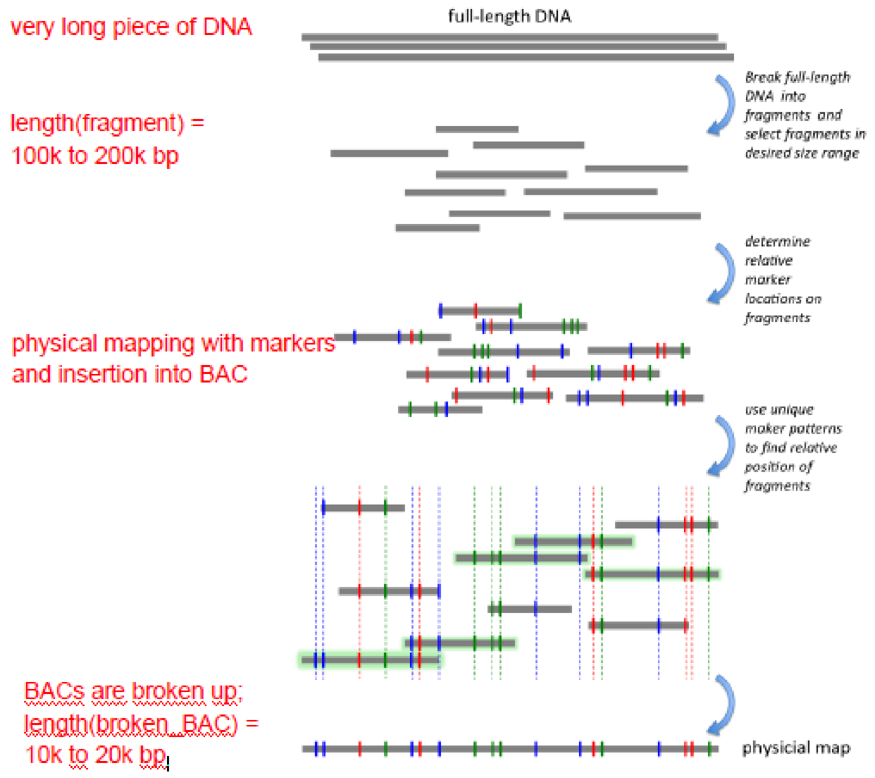
A step-by-step outline of the workflow for sequencing by synthesis:
1) Tens of thousands of copies of the complete DNA to be sequenced are isolated from a specimen.
2) The DNA is mechanically sheared to yield DNA fragments with a typical length of 500-1000bp
3) The DNA fragments are ligated to adapter oligos containing ID-tag, surface attachment and primer-complementary regions.
4) Library molecules are bound to the flow-cell surface so that individual molecules are physically well separated from one another.
5) Individual library molecules are amplified via bridge amplification to generate clusters of identical molecules.
6) 150-300 SBS cycles are performed to determine the sequence of the DNA fragments and any included ID-tag sequences.

3rd generation: sequencing individual molecules: Usage if nanopores. (no need to go into further detail.)

Sequence assembly:

There are two methods: mapbased sequencing (also known as BAC-to-BAC sequencing) and the shotgun method.

Mapbased sequencing:


very long piece of DNA
full-length DNA
Break full-length DNA into fragments and select fragments in desired size range
length(fragment) =
100k to 200k bp
determine relative marker locations on fragments
physical mapping with markers
and insertion into BAC
use unique maker patterns to find relative position of fragments
BACs are broken up;
length(broken_BAC) =
10k to 20k bp
physicial map

Those BAC parts can now be sequenced with primer walking: For this the initial sequencing reaction uses a primer that is complementary to a known portion of the plasmid that is directly adjacent to the 5' end of the inserted fragment. The sequence generated in this reaction will then reveal a substantial stretch of the sequence of the fragment. This sequence can then be used to generate the primer for the next sequencing reaction and so forth. The process is repeated until the appearance of sequence stemming from the plasmid indicates that the entire sequence of the insert has been obtained.

Drawbacks of mapbased sequencing: Large genomes need many BACs (human genome has 6 * 10**9 bp – this corresponds to an absolute minimum of 40k BACs, 600k plasmid fragments and a minimum of 6 million primers). Therefore, it is labor-intensive, a logistically challenging process and expensive.

Shotgun sequencing: …

KEGG: database, good for finding genes involved in metabolic processes such as lactose intolerance. Contains lots of information, drawings, nucleotide sequences, location of genes on chromosomes etc.

SNPedia: large database like Wikipedia.

PubMed: like SNPedia, but it contains lots of papers that need to be downloaded first.

GenBank: contains all the sequences.

NCIB: genomes browsen.

UCSC genome browser: default genome browser for scientists. It is the google maps equivalent for genes, chromosomes etc.

1000 Genomes: distribution of alleles in the population.

HapMap: similar to 1000 genomes.

What is the wild type of lactose intolerance? What is the ancestral allele?
One can have a look at the close ancestry of the human (chimpanzee). They are lactose intolerant, so the hypothesis is that humans are lactose intolerant too.
Go to BLAST and compare chimpanzee with humans and put in the right sequence (use scientific names, not trivial names).

ADDENDUM – Quiz:

Grading function (dt. Bewertungsfunktion) and search algorithm: Grading function calculates a numerical value (score) for the quality of an alignment. The search algorithm looks for an alignment with such a score.

HapMap data can be used to find SNPs that are passed on together as LD blocks (haplotypes).

5.3.2017

On genomics and the techniques it uses to analyze whole genomes

PCR-based genome analysis: If the genome of an individual is already known, then one only needs to analyze whether allele A or allele B is present. Use two primers, one that binds perfectly to allele A and one that binds perfectly to allele B. Carry out 2 PCR reactions and observe product with

gelelectropherosis. PCR product of allele A with primer A should be clearly visible. PCR product of allele B with primer A cannot bind efficiently enough to produce constantly or useful proteins, therefore, there should only be little or nothing of it, because of mismatch. This is the case for individuals homozygous for one the alleles A or B.
If the individual is heterozygous, thus, both PCR products would occur equally, since in each case, one primer will be able to bind correctly to the DNA.

Contemporary version of the PCR-based genome analysis: **fluorescent based TaqMan method**:
4 primers needed, 2 primers are normally used for the PCR, primer A and primer B have a fluorophore (one that radiates green light, the other red light for example) and a quencher molecule at the other end (the quencher molecules supresses the fluorescence of the fluorophore by quickly absorbing the emitted wavelengths). In the first cycle, the normal primers produce either the DNA strand with the allele A or allele B. In the second cycle, during the synthesis of the new strand, either primer A or primer B will bind (the labelled primer binds because of the perfect sequence complementary extremely well to the new DNA strand). DNA polymerase cuts the labelled primer through the naturally happening exonuclease activity in separate nucleotides. Those nucleotides diffuse and the quencher molecule cannot suppress the fluorophore anymore. Through adequate lighting there will be a specific fluorescent signal, whether the gene A or gene B is present. Special real-time PCR machines allow us to observe live, in what concentrations the fluorescent signals occur.

Microarray based genome analysation: Many genotype analysis reactions occur on a small chip, so called SNP chip. It is useful to identify the genotype of chosen loci (is in between PCR reactions and whole genome sequencing). DNA is fragmented and denatures. It can then bind to the SNP chip specifically and one gets fragmented DNA double strands on the chip. The chip contains the correct complementary configuration.

Targeted DNA sequencing in the exome: De novo mutations often have an effective, observable phenotypic effect in the protein coding region of the genome (for example by adding an early stop codon into the DNA sequence). Firstly, one breaks up the DNA into 100 bp long DNA fragments and add oligos to them that have a complementary sequence on them (and the complementary sequence is complementary to the protein coding regions). This allows hybridization (the ends of the oligos are labelled with oligos). Those hybrids can be bound to streptavine labelled beads. Now, we can wash off the non-coding DNA parts. Denaturing the DNA anew will remove them from the beads, resulting into a library of protein coding DNAs. Use sequencing by synthesis to study the DNA fragments.

Protein DNA interactions: Idea behind this is to find out all interactions of a chemical species (such as a transcription factor, a polymerase, a histone etc.) with DNA.

Apply the crosslinking chemical formaldehyde for it produces covalent and reversible bonds between proteins and the DNA in question. The DNA is cut and immobilized on beads. Specific anti bodies are used to remove the protein (with the DNA) from other cell components. Crosslinking is reversed and DNA can now be sequenced.

7.3.2017

Modern methods in genomics:

Questions – Muddiest points (see slides of this week again)

Use google, YouTube and look at the companies' online presence of the respective techniques for nice graphics.

Why would one want to sequence something with targeted DNA sequencing?
Answer. For identification in an unknown probe whether a gene is present and to find new small variations of a gene.

In Chip-Seq: How do you find regions in a gene when the protein is not known?
Answer. -

What happens during the PCR reaction with the allele specific primer in the two mentioned processes?
Answer. In classical PCR, the primer is the template and it is not "eaten".
In TaqMan, we have two reverse and forward primers. We do not care for the PCR product, we only need it as the template so that the Taq primer can bind to it. It is devoured by the exonuclease process of the polymerase. The new template for the Taq primer is the complementary sequence of the template during the initial PCR.

Nucleic acid analysis as primary tool in genomics:
DNA level: sequences and small differences between sequences, structural genomics changes, chemical modifications of DNA.

RNA level: transcript mRNA abundances, various species of RNAs.

ADDENDUM – Quiz:

Microarray and PCR based genotyping methods are often used to find already known polymorphisms that occur often.

11.3.2017

Bacterial genetics

A special type or recombination in DNA is the jumping of genes to other parts of the DNA. This process is called transposition, its enzymes are transposases which are responsible for the movement of the transposons and they work independently of the rest of the DNA. The genes can jump to non-homologous sequences. This process has been especially well examined in bacteria, where the transposon typically looks like a phage or plasmid.

Transposons are in the host cell's genome edited. The transposons DNA possesses the DNA sequence for the necessary tools to relocate in the genome (cutting and editing). One can take advantage of this function to add genes, so that they are carried over as well.

Since too many transpositions would eventually destroy the host cell completely, the transposons have developed very subtle mechanisms so that a transposition occurs every 10**3 to 10**8 depending on the type of the transposition. (A transposon might jump into a gene and inactivate it.)

Bacteria as model systems:

Bacteria are haploid, they have a short generation duration, they reproduce asexually, huge populations can be easily cultivated.

Transposon mutagenesis:

Transposons jump randomly into a gene of a bacterium often leading to its inactivation. That way, one can observe its effects on survival and reproducibility. Since the transposons is known, one can easily figure out the inactivated gene with PCR or NGS-methods. Normally, a transposon contains further marker genes, e.g. the immunity to an antibiotic so that the affected bacteria can be identified more easily.

The following criteria define a good transposon for mutagenesis:

1) Transposition occurs with a high frequency for more mutations.
2) The transposon has a low selectivity, such that all genes are affected about equally.
3) Transposon contains a selectable gene, so that the affected bacteria can be selected more easily.
4) The transposon can induce transposition in many different bacteria kinds, so it can be used and observed in different bacteria.

Methods on how to insert a transposon in a bacterium's genome and how to identify it:
Add the desired gene into the donor bacterium via a plasmid and let it multiply there (often in E.coli). Then let the recipient bacterium grow in the same medium as the donor in order to ensure cell-to-cell contact. Through conjugation, the plasmid will get to the recipient. The recipient should be unable to express the genetic information on the plasmid, so that it can only express the desired gene when through transposons it jumps over to the recipient's genome and is permanently integrated there. In case of a resistance gene, one can add an antibiotic to select for the desired bacteria. Also, remove the donor by changing growth conditions.

ADDENDUM – Quiz: Many molecular processes in bacteria are very similar in eukaryotes which is why bacterial lifeforms can be good model organisms to observe molecular processes. Also, they are haploid, reproduce very quickly, they form large cell populations on limited space etc.

E. Coli's genome is 4.6 million bp long.

Base pair substitutions occur with a frequency of 2*10**-10 per cell division and per base pair. Base pair substitions are 10 times more frequent than indels.

Def. indel: **in**sertion and/or **del**etion.

17.3.2017

Yeast genetics:

Advantages of yeast: short generation duration (only 90 minutes, that is almost as fast as bacteria; human cells take 24h to duplicate under optimal conditions);
experiments can be done in either haploid or diploid stage and growth occurs in both stages;
plasmids can be very easily added to the yeast and replicated as episomes there (since the origin of replication in yeast are known);
two haploid yeast cells can merge together into a diploid cell and one can observe dominant or recessive alleles;
a diploid cell can give rise to 4 haploid cells through meiosis which allows the observation of different

phenotypes of multiple mutants (those 4 cells are called ascospores and they are held together in a "sack");
a single cell can be isolated and incubated and through asexual reproduction (budding) one can observe colonies of cells with identical genetic background (the "offspring" is considered to be a clone from the original single cell) (important to identify mutants and the loci of the mutation in the genome);
small compact genome

Def. episome: A piece of DNA that can exist and replicate autonomously in the cytoplasm or on a chromosome (mainly found in bacteria, but also in yeast).

Facts about yeast:

G1-phase: no budding, chromosomes are diffuse
S-phase: budding, sister chromatids are created (duplication of chromosomes)
G2-phase: budding, cell nucleus is next to the bud, sister chromatids still diffuse
Mitosis: chromosomes are segregated between mother and sister cell. In G1-phase, the cell nuclei are distributed through cytokinesis

Yeast has 16 chromosomes, 6600 genes and about 12 million base pairs. It does not have many introns, genes are rather short in length and therefore gene density is quite high. It also possesses a mitochondrial genome and the 2 mu plasmid. Genetic redundancy is low in yeast.

Def. Tetrade: Sack of 4 haploid cells (spores). This form occurs when environmental conditions are not in favour of the yeast cell. A tetrade consumes less energy and is more resistant to unfavourable environmental conditions. The cell wall is very thick of the sack (ascus) and protects the tetrade.

Mutagenesis methods:

Def. homologous recombination: A type of genetic recombination in which nucleotide sequences are exchanged between two similar or identical molecules of DNA. It is most widely used by cells to accurately repair harmful breaks that occur on both strands of DNA, known as double-strand breaks.

A chemically-induced mutation can be achieved with ethylmethansulfonate (EMS). EMS is used in temperature-sensitive mutations, since it methylates the DNA bases which results into the DNA polymerase adding the wrong base. It is often used for point mutations.
Intense radiation of UV-light leads to transitions and transversions.

We call such cells, conditional mutants that can only live under certain circumstances. It is especially helpful in diploid organisms in order to observe their mutations and phenotypes.
Heat sensitive mutations for example influence the stability of proteins. Cold sensitive mutations disturb protein interaction so that their respective reactions do not occur or are hindered.

Def. suppression analysis: Organism has a mutation. A second mutation is introduced or the gene activity of another gene is amplified in order to recover the initial phenotype. The mutated phenotype is no longer observed.

Def. synthetic lethality: Organism has a mutation. A second mutation is introduced or the gene activity of another gene is amplified so that the organism can no longer live under the current conditions. Organisms therefore die and this method is applied in order to observe protein interactions and signalling pathways.

Kor.: Synthetic lethality can identify redundancies in cell. It can also identify partial loss of functions, when it is a linear pathway.

ADDENDUM – Quiz:

Yeast also have very similar molecular and cellular processes like eukaryotes. Also, they are haploid, but they can also exist in a diploid stage. In order to observe lethal mutations, one mutates yeast in such a way, that their lethal mutation is only lethal under certain circumstances, such as temperature for example.

21.3.2017

In yeast, there are about 6000 genes of which 15% of all expressed genes have an unknown function and cellular purpose in yeast. The removal of those genes does not yield to any definite conclusion of their function, because no clear phenotype can be observed in yeast.

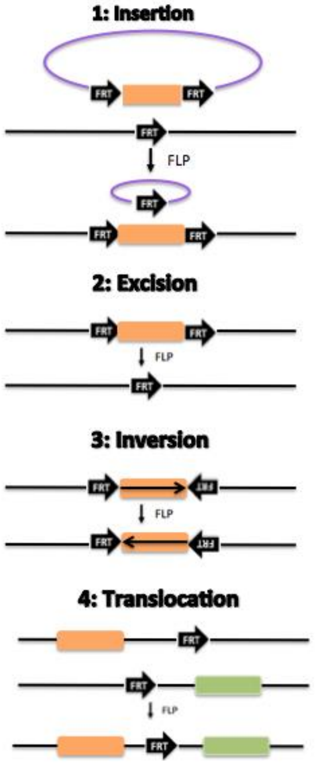
Plasmids can be very easily added and used in yeast.

Methods: complementation and sequencing

In sequencing, one simply sequences the mutant's genome and finds the mutations and compares it to the wild type. easy and cheap.

In complementation, one transforms the plasmids. yeast can very easily take up extra-chromosomal DNA. One checks whether a plasmid can save the yeast cell in its mutation. This complements the mutated phenotype and restores the phenotype. Then, simply sequence the plasmid which is even easier.

27.3.2017



On drosophila:

Facts: 4 chromosomes (3 autosomes and 1 sex chromosome) and about 180 million bp. Females have a segmented coloring of their abdomen, while males have their abdomen completely black. 15000 genes, 50% of them are homologous to humans. It is often used as a model animal for higher order animals such as humans. Also, it is often applied in embryonal research and its phenotype can often be observed by the naked eye or with a microscope.

Def. balancer chromosome: A chromosome with 3 criteria: It contains a lethal recessive mutation when it is homozygous, they possess one or more inverted DNA parts which lead to deletion or duplication during meiosis if crossing-over occurs, they possess a dominant phenotype marker.
Balancer chromosomes are used in model organisms in order to select the wanted organisms more easily and in order to preserve an otherwise lethal mutation (which is on the mutated chromosome) in a species. Ex: Drosophila.

EMS is an extremely effective mutagen that typically leads to transitions (point mutations).

Def. Flp-FRT recombination system Analogous to the Cre-loxP system. FRT (flippase recognition targets) sites and the flippase transgene have to be introduced artificially into most organisms. It is applied during mitosis to produce specific mutations.

Clonal screens vs. F2-screens: Advantage clonal screens: No need for a 2nd generation. Recessive phenotypes can be observed in F1 generation (homozygous mutants are produced in the F1 generation through homologues recombination in mitosis). Also, one can induce tissue-specific mutations or control mutations at a certain stage in development with Flp recombinase (Flp-FRT recombination system or Cre-loxP system or CRISPR/Cas9). Since a homozygous embryo is produced at a later stage in its development, one can omit in specific situations a lethal homozygous mutation.

8.4.2017

RNAi and CRISPR/Cas9

Def. Watson-Crick base pair rule: Ratio of A-T and G-C is 1:1. A pairs with T (pair with U in tRNA) and G pairs with C. There are alternative pairings that might occur in tRNA called reverse pairing, Hoegsteen pairing and Wobble pairing.

There are different types of RNA-interference, such as miRNA, piwiRNA and siRNA.

On micro RNA (miRNA): Its information is encoded in the introns of DNA and the genetic information is sometimes found in close neighborhood of other genes. miRNA does not code for a protein, it is used to suppress or to negatively regulate the expression of one or more mRNA. miRNA has complementary sequences according to the Watson-Crick base pair rule and can bind to the mRNA. Since it is not 100% specific, it might also bind to other mRNAs and negatively regulate their protein expression.

How miRNA is made: The sequences is first transcribed into a primary transcript, the pri-miRNA. Then, splicing removes the introns, a polyadenyl chain is added and a 5' cap is added to the other end.
The pri-miRNA sequence contains inverted repeats that can hybridize and form hair pin structures. A protein complex consisting of Drosha and DCGR8, called a microprocessor, recognizes the pri-mRNA. Drosha cuts off the miRNA-stem-loop out of the pri-miRNA and isolates thus the pre-miRNA (60-80 bp long).
Pre-miRNA is exported into the cytoplasm with exportin 5. There the dicer protein recognizes double-stranded pre-miRNA and cuts it in such a way that we finally get miRNA-duplex (double stranded miRNA). Lastly, unneeded sequences are removed such that the mature single stranded miRNA is in the cell. The miRNA can now bind to the RISC-complex (RNA induced silencing complex).
The RISC-miRNA-complex can regulate protein expression in two ways now. If the miRNA is perfectly complementary to the mRNA, the complex will recognize the mRNA, bind to it and cut it in half. The cell will recognize the destroyed mRNA as non-functional and degrade it.
If the miRNA is not perfectly complementary to the mRNA, the RISC-miRNA-complex will still bind to the mRNA and negatively regulate its protein expression (translational block). The efficiency of translation is reduced. In such a case, a miRNA can influence several hundred genes.

On small interfering RNA (siRNA): siRNA is always perfectly complementary to the targeted mRNA. It is always used to destroy the mRNA (plants make use of siRNA to protect themselves from foreign mRNA). siRNA is commonly used for gene-knockdown experiments, where the RNA is rendered inactive instead

of the DNA (such is the case in a gene-knockout experiment). There are two common ways: One can either synthesize the siRNA chemically and add it to the cell or write the corresponding DNA and add it to a vector. This vector will be introduced into a cell that can translate the genetic information on the vector. The DNA for the siRNA will contain inverted repeats such that it can form short hairpins. We call the transcribed DNA shRNA. The dicer enzyme will cut the double-stranded shRNA into siRNA.
The chemical approach will only knockdown the target RNA for some time, since it is degraded by the cell. On the other hand, the vector approach will always generate new siRNA, since the genetic information is transcribed more than enough.

How can one design such a siRNA sequence: As a general rule, it should be specific to the first 50-100 bp of the targeted mRNA, have a CG-percentage of 50%, start with two A's and internal hybridizations should be impossible.
Since a siRNA might also partially bind to another mRNA and negatively regulate the protein expression similar to miRNAs, off-targets effects (unwanted phenotypes) can occur.
Firstly, one can make use of bioinformatical methods to choose a siRNA specific enough. Secondly, 3 experiments with independently different siRNAs will be conducted and they all have the exact same target mRNA. If all 3 experiments show the same phenotype, it is unlikely that an off-target effect has been observed. In a follow-up control experiment, one uses one of the siRNAs to knockdown another gene in order to confirm that the phenotype is not the result of an injection of an siRNA (Ex.: through a non-specific immune reaction).

On CRISPR/Cas9: The system has its origin in bacteria where it functions as a bacterial immune system. The bacterial genome possesses clustered regularly interspace short palindromic repeats (CRISPR) parts. They are made of short repeat sequences (20-40 bp in length) with spacer sequences in between. The spacer sequences contain the genetic information of a viral genome or other harmful DNA for example. After translation, the pre-crRNA is cut into smaller pieces called crRNA. Each of them possess one segment of the repeat sequence and one segment of the spacer sequence.
The crRNA binds to a tracrRNA (trans-crRNA). The result is a heterodimeric double stranded RNA.
This heterodimeric RNA binds to a multi-domain protein called Cas9. This huge complex is now able to recognize invasive harmful DNA and destroy it.

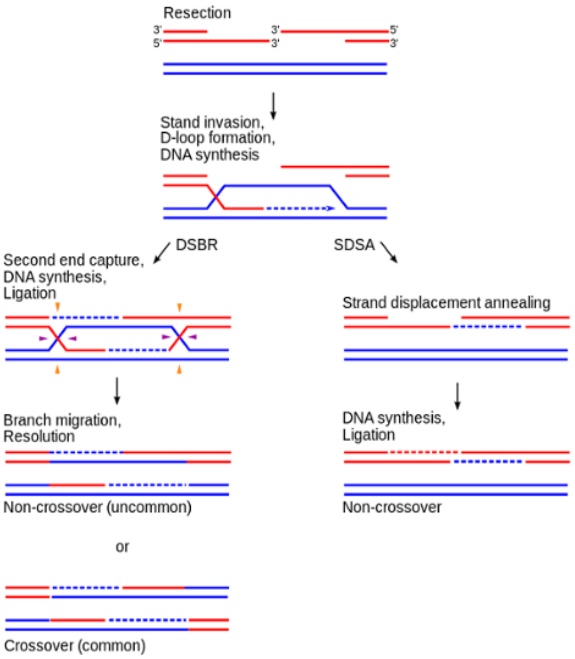
The Cas9 protein complex unwinds the target DNA and the spacer crRNA part can hybridize to the complementary DNA strand, if there is next to the foreign DNA a PAM sequence (protospacer adjacent motif, typically 2-6 bp long, in the case of Cas9 it is NGG). This DNA, registered by the spacer-sequence as harmful, is destroyed.

Def. single guide RNA (sgRNA): A long continuous RNA consisting of crRNA and tracrRNA.

How to do it in the lab: It is incredibly straightforward if the genome is known. First, one searches a PAM sequence (in S. pyogenes it is NGG, which occurs relatively often). The first 20 bp (from the 5' end) are then cloned in single guide RNA, which is in an expression vector with the right promoters already. This defines our spacer sequence. Another vector responsible for the expression of the Cas9 protein is inserted into a cell together with the first expression vector.

Def. non-homologous end joining (NHEJ): After a double strand break, the broken ends are simply joined together without a template. If it occurs in the protein coding region of the genome it will lead to the destruction of the protein (deletion/loss-of-function).

Def. homology directed pair (HDR): Uses the homologous template DNA for repair (has to be intact).



Def. oncogene: A gene that is capable of causing cancer. Such genes are expressed in high rates in mutated tumor cells.

QUESTIONS:

1. **Using CRISPR/Cas9 for germ line gene editing. What are the benefits and what are the risks?**

   Benefits: Curing genetic diseases, but only easy for monogenetic diseases, since diseases involving several genes, it will be very hard to understand all these networks as of now.
   Easier to make KO mice etc. with CRISPR/Cas9 to study them (already carried out in research)

   Risks: Random damages to the genome can occur; unknown consequences might lead to new problems. It is much safer to choose an embryo that is heterozygous to a genetic disease, only makes sense for homozygous embryos that are really rare.
   Is it ethical to produce "superhumans" while some people cannot have access to improve their genome -> could lead to an unfair system.
   Introducing new genes might cause problems for coming generations and offspring might be unhappy about the parent's decision.

2. **Using CRISPR/Cas9 to generate gene drives: What are the benefits and what are the risks?**

   Benefits: Curing illnesses (e.g. Malaria), but maybe the illness itself might evolve and become resistant.

   Risks: It is not known, what gene drive might do to ecosystems.

3. **Using CRISPR/Cas9 to generate GMOs for agriculture. What are the benefits and what are the risks?**

   Using CRISPR/Cas9 it will be impossible to distinguish between wild type organisms and CRISPR/Cas9 organisms.

ADDENDUM – Quiz:

The CRISPR/Cas9 system can be used to insert a sequence in a genome through HDR (homology directed repair). It is also used to create targeted mutations in the cellular genome. Since a double strand break is often not perfectly repaired, but contains small insertions, the result is often an insertion mutation.

The natural function of CRISPR/Cas9 is to save detrimental DNA in the bacteria's genome and to destroy such DNA when it is found in the cell (such DNA originates from viruses).

The RNAi system can be used to permanently KO gene activity or for a certain period of time.

24.04.2017

Cell line genetics

Def. primary cell culture: A cell culture that originated from a donor body. Such cell cultures often need very specific environmental predispositions in order to proliferate, such as growth factors, energy sources, composition of the surface (it has to simulate an extracellular matrix so that a cell can attach it to it). They cannot divide indefinitely.

Def. stable cell culture: A cell culture that has certain genetic mutations in order to divide indefinitely regardless of the environmental predispositions. They have been immortalized.

Def. immortalization: A process that enables cell cultures to proliferate regardless of their environment and cell cycle ⇔ the cell's cell cycle has been manipulated or deactivated, such that it never stops dividing. Ex. Cancer cells show this uncontrollable division phenomena.

There are two ways to make artificial "immortal" cells: Firstly, one can fuse a somatic cell with a cancer cell (we call it a hybridoma cell). Secondly, there are certain viral genes that can be expressed in a somatic cell in order to avoid senescence.

Since animal somatic cells are diploid, recessive phenotypes cannot be observed that easily, because somatic cells reproduce asexually. In asexual reproduction, there is no mating and no crossing-over respectively. Also, the animal genome has redundant genes.

Forward genetics ⇔ start with a phenotype and find the genotype (classical approach; useful in simple model organisms).

Reverse genetics ⇔ start with genotype and mutate it, thus find the phenotype (often carried out in cell cultures).

RNAi, CRISPR/Cas9 and the artificial introduction of genes via plasmids in a cell are used to induce overexpression of genes. Marker genes are also introduced to the genome along the other genes to verify its successful introduction.

The HPA1 cell culture originates from leukemia patients. This cell culture is haploid in all chromosomes except for chromosome 8 and 15. These somatic cells are mostly haploid and recessive phenotypes can be observed a lot easier. Of course, the HPA1 cell culture also has other mutations, so that it is possible for such cells to survive in the haploid state in the first place. They are typically smaller in size compared to their diploid version. One has to keep in mind, that even if they are practical to work with, it is not the somatic cell's normal state of operation.

25.04.2017

Phenotypic plasticity vs. karyotypic plasticity

There is no correlation between how the chromosomes look like and the phenotype.

There is no definite reason to have 2 copies of each chromosome. Of course, one chromosome can still be used as a backup if a mutation occurs on the other one, but on the other hand, having 4 copies per chromosome would be even better by this logic.

Mutations occur within the body all the time, but we do not notice them as we have a second functional chromosome. It becomes more evident in reproduction (gametes with mutations on them).

There are aquatic lifeforms that can switch between haploid and diploid stages of life. The advantage in haploid state is that it needs less nutrients and proteins to live, since a haploid genome simply needs less building blocks than a diploid genome.

In mosses, one can observe a clever trick why the sporophyte is diploid. Instead of letting the reproduction depend on phenotypic variables, the genotype is more important for reproduction.

Mutations in haploid genomes affect the organism a lot stronger regarding its natural selection (both in a good and a bad way) than in a diploid organism which still have an extra unmutated copy in the case of heterozygous organisms.

Nowadays, natural selection does not act on metabolic changes anymore, but more on the metalevel of genetic information, such as cell-to-cell interaction, organization of organs and so on.

29.4.2017

Genome wide association studies/scans (=: GWAS)

Quantitive phenotypes such as blood pressure or cancer risk are controlled by several genes. Each genes contributes only a little to the quantitive phenotype, which makes it quite difficult to identify the underlying genotype. Classical mapping methods are difficult to apply to quantitive phenotypes as we know it from model organisms.

Method: (candidate gene association study): Assume a gene that might be mechanistically involved in a phenotype. Then, observe its genetic variations for associations.
Ex. Gene: INSR gene for insulin study. Make KO phenotypes etc. to see if phenotype changes or not.

An illness is rare ⇔ less than 1 in 2000 persons are affected.
A rare illness often has a mutation in a risk gene, such that a specific enzyme becomes dysfunctional or is underperforming. By definition, only a few people are affected, although these people may have a SNP mutation in different parts of the gene, damaging the enzyme unequally.

The opposite are frequent diseases. Hypothesis: common variant-common disease hypothesis.
Method: GWAS are designed to statistically evaluate genotypes and the associated phenotype.
Note: A priori, no genes are assumed. Starting point of a GWAS is the phenotype.

Assumption in the data analysis of a GWAS: T follows a normal distribution (sometimes it's easier to use the log-function on every data point to get normal distribution). Also, the variance of all data sets should be the same, else, misleading significances may occur.

Note that not only does a genotype influence the phenotype but also other covariables such as gender, the participant's nutrition, concentration of a molecule in an organ etc. Also, we get noise data sometimes.

For statistical significance, p = 0.01. But since in GWAS, there are often more than a million tests, we expect statistically that some tests will already be >p. Therefore, we apply the Bonferroni-correction which is: p/[NUMBER_OF_TESTS] = 10**-8 for example. The Bonferroni-correction also assumes that all these tests are independent, but since they are correlated with the linkage disequilibrium, we use p = 10**-7.5. When one test is significant, p is modified to p' = p/2 and so on for the next test.

Def. linkage disequilibrium: Blocks of alleles that are always transferred together. The non-random association of alleles at different loci in a given population. Loci are said to be in linkage disequilibrium when the frequency of association of their different alleles is higher or lower than what would be expected if the loci were independent and associated randomly.

Def. haplotype (as in haploid genotype): A group of genes that is inherited together from one parent.

ADDENDUM – Quiz:

The length of a microarray in such a study is about 500k to 1000k bp long.

Due to SNPs being on LD blocks, these SNPs all have the same p-value and are associated with the same phenotype. Therefore, one cannot determine if a SNP is causal or if the SNP has been passed on with a causal SNP in the same LD block.

6.5.2017

Epigenetics

All cells in an organism have the same genome sequence. Why are there so many distinct types of cells? The mother cell does not only pass on its genetic information to the daughter cells, but also the expression patterns. Those are often reversible chemical modifications that influence the expression. Such effects lead to the diversity of cells found in an organism.

This effect is not only subject to mitosis but also to meiosis. The offspring also inherits epigenetic modifications (transgenerational epigenetic heritage).

Epigenetic processes occur parallel to Darwinian evolution (mutations and natural selection). Within only one generation, a population can adapt to environmental hardships.

Molecular mechanisms that regulate DNA expression

Chromatin density: The denser the DNA is coiled around the histones the harder it is for the DNA-polymerase to get to the DNA regions. Very dense and hardly accessible regions are called heterochromatin and are therefore not expressed. Easily accessible regions are called euchromatin and are accessed often and easily.
A process that modifies the density is called chromatin remodeling.

Histone modification: The central structural element of chromatin is the nucleosome. The core of the nucleosome is made up of an octamer of 4 histone proteins (H2A, H2B, H3, H4; two of each). This core is always well-defined. A second structure called **histone tail** is attached to it. It is more flexible which can be influenced by chemical modifications such as:
1) modification of mono-, di-, tri-methylation of Lys
2) methylation of Arg
3) phosphorylation of Ser, Thr, Tyr
4) acetylation of Lys and ubiquitation and SUMOylation

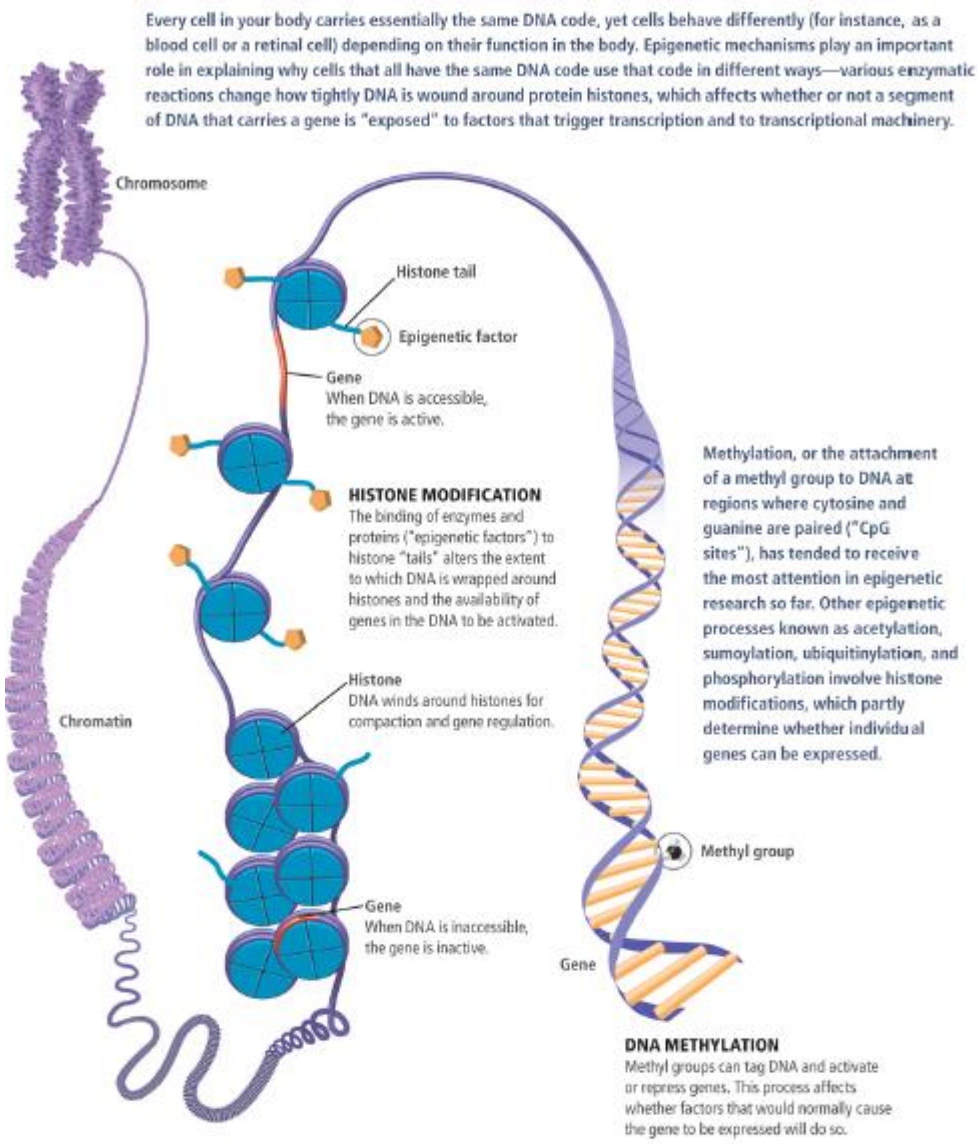
Histone tails are often modified to change its chemical and physical properties that influence gene expression:

Histone modifications occur locally. There can be several active and inactive DNA parts right next to each other. Exception: If the X-chromosome is inactivated, the whole chromosome is packed densely and is rendered inactive.

Histone modifications and its effect on DNA is reversible. There are specific enzymes that control these modifications (called readers and writers). These enzymes are controlled by the cell.

Acetylation of lysine side chains changes its electrostatic properties (normally, it has a positive charge, but then it becomes neutral). This influences the strength of the bond between histones and DNA. This weakens the association which results into less dense packages that are therefore more easily accessible for transcription.

The underlying cause for methylated lysine side chains seems not to be of physico-chemical nature, but on the grade of methylation. It is postulated that there are specific recognition proteins that will modify transcription according to the grade of methylation (mono, di, tri).



Every cell in your body carries essentially the same DNA code, yet cells behave differently (for instance, as a blood cell or a retinal cell) depending on their function in the body. Epigenetic mechanisms play an important role in explaining why cells that all have the same DNA code use that code in different ways—various enzymatic reactions change how tightly DNA is wound around protein histones, which affects whether or not a segment of DNA that carries a gene is "exposed" to factors that trigger transcription and to transcriptional machinery.

Gene
When DNA is accessible, the gene is active.

HISTONE MODIFICATION
The binding of enzymes and proteins ("epigenetic factors") to histone "tails" alters the extent to which DNA is wrapped around histones and the availability of genes in the DNA to be activated.

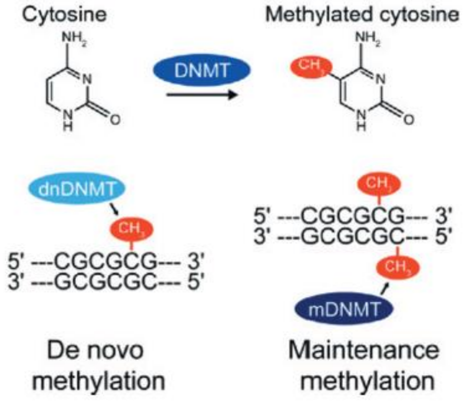
Histone
DNA winds around histones for compaction and gene regulation.

Gene
When DNA is inaccessible, the gene is inactive.

Methylation, or the attachment of a methyl group to DNA at regions where cytosine and guanine are paired ("CpG sites"), has tended to receive the most attention in epigenetic research so far. Other epigenetic processes known as acetylation, sumoylation, ubiquitinylation, and phosphorylation involve histone modifications, which partly determine whether individu al genes can be expressed.

DNA METHYLATION
Methyl groups can tag DNA and activate or repress genes. This process affects whether factors that would normally cause the gene to be expressed will do so.

DNA methylation: DNA methylation is a very central process. A methyl group is normally added to the 5' position of a cytosine base from a methyl donor source aided by DNMTs (enzymes).
De novo DNMTs are able to add a methyl group to a cytosine base that possessed none previously.
Maintenance DNMTs add methyl groups to a daughter cytosine base if the template strand had one.

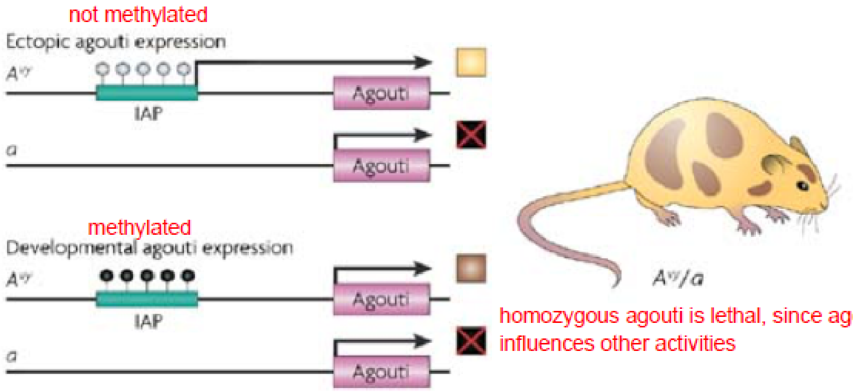


Such a methylated cytosine base will have a guanine base next. We call this occurrence CpG (p stands for phosphate). So, a methylated cytosine base will always have a guanine base next. CpGs are unequally distributed. There are regions with a very high CpG density, we call these regions **Cpg islands**. Typical regions are promoters or retrotransposon sequences.

Def. differentially methylated regions: Not all cell types have the same amount of CpG islands. CpG island occurrences differ in promoters. On the other hand, all retrotransposon sequences are equally methylated in all cells.

Def. transcriptional silencing: Methylation of CpG dinucleotides in regulatory parts of the DNA will lead to a weakened transcription depending on the degree of methylation.

On Agouti mice as an example for epigenetics:



not methylated

methylated

homozygous agouti is lethal, since agouti also influences other activities

the cryptic promoter is constitutively active (not dependent on hair growth unlike the normal promoter)

Abbildung 6 Schematische Darstellung des Agouti Locus in einer A^{vy}/a Maus. *Im a Allel (jeweils unten) ist das Agouti Gen zwar unter der Kontrolle des „natürlichen“ Agouti Promoters, aber das Transkript ist inaktiv und eine a/a Maus würde deshalb eine dunkelbraune bis schwarze Fellfarbe (quadratische Box) haben. Im A^{vy} Allel kann die Expression des Agouti Gens von einem, sich in der benachbarten retrotransposonalen IAP Sequenz befindlichen, kryptischen Promotor angetrieben werden. Sind die potentiellen DNA Methylierungspositionen, wie im oberen Teil der Abbildung gezeigt, nicht methyliert (weisse Kreise) ist dieser Promotor konstitutiv aktiviert. Das Agouti Gen ist also permanent exprimiert und das produzierte Fell ist gleichmässig gelb. Ist die IAP Sequenz, wie im unteren Teil gezeigt, methyliert (schwarze Kreise) ist der kryptische Promotor inaktiv. Dadurch steht das Agouti Gen wieder unter der Kontrolle seines eigenen Promotors, die Expression des Agouti Signal Proteins ist also mit dem Wachstumszyklus der Haarfollikel koordiniert und die Haare haben die wild-typartigen dunklen Spitzen mit dem gelben Ring, welche die klassische hellbraune wildtyp Fellfärbung ergibt. Das Fell einer A^{vy}/a Maus ist oft ein Mosaik aus wildtypfarbigen Bereichen in denen die IAP Sequenz methyliert ist und gelb gefärbten Bereichen in denen die IAP Sequenz demethyliert ist.*

Def. parental conflict hypothesis: A conflict between mother and its offspring: Paternal genetic information is interested in asserting itself in the mother (fitness of paternal genes), thus it is often growth promoting. Paternal gene’s fitness is increased theoretically when it can use up as many resources from the mother as possible. Since the mother has to balance her own survival and fitness with the nourishment of her offspring (during pregnancy), thus the maternally expressed genes tend to be rather growth limiting.

Selfish gene view: Gene centered view of evolution. Genes that provide the organism with completive survival advantage will be selected for by evolution.

Def. coadaption theory: Imprinted genes act adaptively to optimize felt development as well as maternal provisioning and nurturing.

13.5.2017

Cancer genomics

Types of cancer: carcinomas (derived from epithelial cell – around 80-90% of all tumors), sarcomas (derived from muscle, bone or connective tissue cells), leucomas (derived from hematopoietic (blood) cells), neurological tumors, germ cells lead to germinoma.

What causes cancer?

External factors (originating from the environment): ionizing radiation (UV-light, radioactive radiation), DNA viruses (papilloma, HepB, EBV), retroviruses (HTKV1, HIV), bacteria (heliobacter pylori), cancerogenic substances (aflatoxin, benzpyren).

Internal factors: mutations in either protooncogenic genes (promote cell growth and prohibit cell death) and tumorsuppressor genes (slow cell growth down, induce cell death (apoptosis), when necessary).

Def. protooncogenic genes: Genes, who promote cell growth and prohibit cell death. A mutation in such a gene will lead to a gain-of-function, which amplifies cell growth and/or makes mutated cells immortal. They are then called oncogenes
Ex.: • c-Ha-Ras, c-Ki-ras, c-N-Ras (GTPasen in Zellproliferationspathways)

• c-Raf (Serin/Threoninkinase)

• c-Jun, c-Fos, c-Myc (Transkriptionsfaktoren)

• c-Src (cytoplasmische Tyrosinkinase)

• c-Sis (Wachstumsfaktoren)

• c-ErbB, HER2 (Wachstumsfaktorrezeptoren)

• BCl-2 (Apoptose Inhibitor)

Def. tumorsuppressor genes: Genes, who slow down cell growth and/or induce apoptosis when necessary. Mutations in tumorsuppressor genes lead to a loss-of-function such that those mutated cells will grow faster and damage the neighboring controlled cells. These mutations are normally recessive, which is why the mutation has to occur on both chromosomes.
Ex.: • RB (Verlust der Funktion führt zum Retinoblastom in der Netzhaut des Auges)

• P53 („guardian of the genome”, Transkriptionsfaktor, Kontrolle des Zellzyklus, Apoptose, DNAReparatur nach DNA-Schädigung)

• PTEN (Protein- und Lipid Phosphatase im PI3K/Akt Signalweg)

• Smad4 (TGF-beta Signalweg)

• Tsc1 und Tsc2 (Hemmung des mTOR Signalwegs)

• APC (Kontrolle des Wnt Signalwegs)

Def. driver mutation: A mutation that gives a selective advantage to a clone in its microenvironment, through either increasing its survival or reproduction. Driver mutations tend to cause clonal expansions.

Def. passenger mutation: A mutation that has no effect on the fitness of a clone but may be associated with a clonal expansion because it occurs in the same genome with a driver mutation. Passenger mutations can also be deleterious for cancer cells.

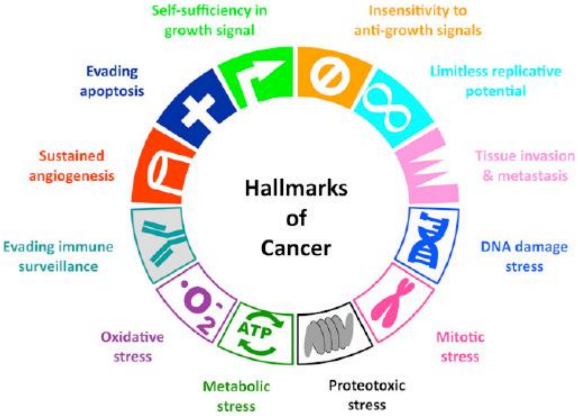
In our current model, tumors emerge from a once healthy cell (called the most common ancestor) in which some driver mutations occurred. Such a cell will keep replicating resulting into more cells with the same driver mutations. In these cells, different driver mutations can occur, leading to a competition within the microenvironment. Often, driver mutations promote more mutations until eventually these cells are classified as tumor cells.

Classification: 6 criteria

1) Sustained proliferative signaling & independent self-supply
2) Evading growth suppressors
3) Resisting programmed cell death (apoptosis)
4) Enabling replicative immortality (through extreme telomere failure of function)
5) Inducing angiogenesis in order to supply cancer cells with nutrients
6) The ability for invasion & metastasis

4 further criteria:

7) Genome instability and specific mutations
8) Tumor promoting inflammations in its environment
9) Deregulating cellular energetics (Ex.: changing for aerobic metabolism to anaerobic metabolism)
10) Avoid immune destruction


Self-sufficiency in growth signal
Insensitivity to anti-growth signals
Limitless replicative potential
Tissue invasion & metastasis
DNA damage stress
Mitotic stress
Proteotoxic stress
Metabolic stress
Oxidative stress
Evading immune surveillance
Sustained angiogenesis
Evading apoptosis
Hallmarks of Cancer
ATP
O2

In classical chemotherapy, it is aimed to destroy the fast-growing cancer cells. It also targets other fast-growing cells such as hair follicles, blood cells etc., which is why chemotherapy patients often live with reduced health. At the same time, it blocks the effectiveness of chemotherapies, since excessive medication will destroy the patient.

An alternative are targeted therapies. They require the knowledge of the cancer genome in order to design medicaments that target very specific mutations in pathways that are only present in cancer cells.

Ex.: Vemurafenib (PLX_4032): Used for targeted cancer therapy.

Ex.: Erbitux: Used for targeted cancer therapy. Blocks EGFR. Not effective in all tissues such as lungs.

16.5.2017

The intestinal epithelium has the fastest turnover rate and is often used as a model organism for homeostasis (only 5 days).

Def. stem cell (short): A cell that is able to self-renew and can differentiate into the different cell types of the respective tissue.

How can we find out experimentally what cell is a stem cell?
Lineage tracing by labelling them. After a short period during tracing, stem cells, progenitor and will increase, differentiated cells will lessen. After a longer period, there will only be stem cells expanded over the whole crypt.

One can also use the cre-loxP system to identify adult stem cells.

20.5.2017

Chemical Genetics

Def. small molecule: A small molecule has a molecular weight < 500 Da. In a cell, small molecules are fundamental for our understandings of the mechanisms within a cell. They can regulate pH values, chemoreceptors, work as second messengers and so on. In the context of genetics, small molecules also regulate around 500 of 20000 genes.

In chemical genetics, we observe biological systems with the help of such small molecules. Small molecules can influence the function of gene products and one can observe their role through a changing phenotype.

Chemical genetics are especially effective in higher organisms, since one can observe phenotypes on a cellular and organism level in a really short time in vivo.

Def. forward chemical genetics: Add small molecules to the organisms (bacteria on plates for example) and observe changing phenotype. Discover regulated protein activity and production by the specific small molecule.

Def. reverse chemical genetics: Screen a protein with multiple small molecules to find ligands. Insert protein and the corresponding ligand in a cell to observe the phenotype (since the ligand binds to the protein in question, it will hopefully alter the behavior and activity of the protein in the living cell which results in a different phenotype).

Reverse genetics in general is an ideal tool to understand diseases on a molecular level.

Knocking out essential genes leads to lethality in normal genetics, because whole pathways and networks are lethally disturbed. In chemical genetics, influencing only a single protein with exogene ligands often leads to viable knock-downs. Also, depending on the concentration of the ligands (small molecules), one can observe different grades of phenotypes.

Principally, every gene can be manipulated. The biggest disadvantage of chemical genetics is the low number of known ligands to its proteins. In future, the diversity of small molecules (ligands) will be larger by combining high-through put screenings with the design of structurally diverse components.

Small molecule microarray: An array with around 10k small molecules. Add protein to it and find those who bound covalently to the small molecule.

Cytoblot: Add a whole cell to small molecules. This method allows to observe cell physiological effects, which is not possible with microarrays.

Automated cell imaging: -

Target identification: -

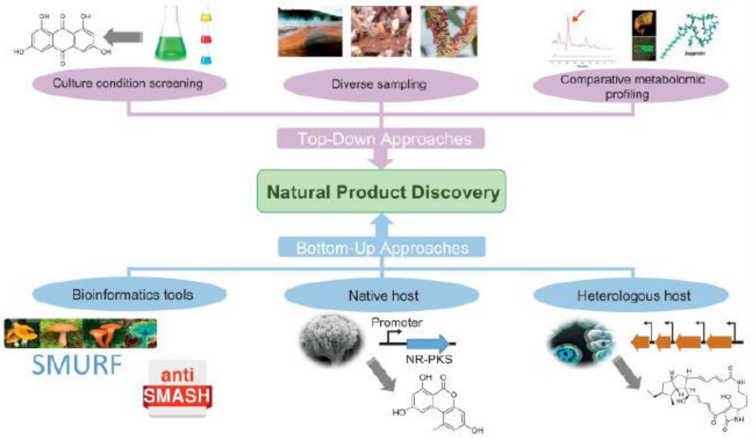
Metabolic mining: Natural products (small molecules) from microorganisms and plants are an excellent starting point for the creation of new pharmaceutical products.

top down: Start with organism and stimulate the production of natural substances without prior knowledge of enzymes or genes. Organisms are being screened under different growth conditions to simulate stress.

bottom-up: Start with genome – if a natural product is present in a cell then logically, the genetic information must be present as well. All the necessary genes for the production lie clustered in the genome. Bottom-up studies are made within the organism, so that the natural product can be observed in its normal environment. Oftentimes, KO studies are conducted, but the genes are inactive under laboratory conditions. That’s why one makes use of Knock-in strategies, that is, exchange the promoter before the desired gene cluster.
The genetic information codes for enzymes that can build the natural molecule within the cell. It often requires many enzymes for the production.

Def. gene cluster (chemical genetics): In chemical genetics, plant cells and fungi are capable of producing natural products (small molecules) within themselves. Every single enzyme is encoded in the genome of the cell and they are placed in sequence. Also, the cell possesses the genetic information for anti-toxic molecules that protect the cell for potentially toxic intermediates.

Thanks to our good understanding regarding natural products and its underlying gene cluster architecture, one is able to make predictions in silico regarding the production of already known and unknown natural substances of an organism.



ADDENDUM:

(LEARN ABOUT TITRATIONKALORIMETRY, MASS SPECTROMY, AFFINITY CHROMATOGRAPHY AND FLUORESENCE POLARISATION AND WHEN THEY CAN BE USED ETC.)

When is chemical genetics superior to classical genetics?
(ANSWER IT YOURSELF.)

23.5.2017

(GO THROUGH THE SLIDES AGAIN, MAYBE TO COPY INFORMAION – SLIDE: 13_CHEMICAL GENETICS SEE SLIDE 17-20 FOR AN EXAMPLE ON FORWARD CHEMICAL GENETICS AND AFTERWARDS FOLLOWS AN EXAMPLE ON REVERSE CHEMICAL GENETICS)

MUDDIEST POINTS:

Difference between RNA-i and small molecules: One aspect are the main building blocks of both. Central difference: RNAi only have one target while small molecules can target more.

Why are there gene clusters? Horizontal gene transfer: good for the adaption of changing environment – when genes are clustered, they stay together during transfer. (?)
Coregulation: one operon can have a whole sequence of genes; therefore, transcription of all necessary genes occurs at the same time.

Why do we find this few active components in the laboratory? Bacteria normally don't live isolated as a single culture – in nature, there are many different participants present with the bacteria in question. Most active components are defense mechanisms, which are not activated in the lab, since they are normally observed in isolation. There is effort to simulate natural environments in order to observe bacteria and the production of their active components.

It is also possible that a bigger small molecule might not enter the cell, when it is outside of it (like in the wells for example) due to the bias of cell membrane permeability.

25.5.2017

Metagenomics

Def. metagenome: The genomic information of a community of organisms (all genomes of every organism that is part of the network/community).

Kor.: 1) Metaproteome ⇔ all proteins produced and used by a community either by the organism itself or by other organisms.
2) Metatranscriptome ⇔ all mRNAs that are transcribed in such a community (and analyzed).

Such a network is not easily simulated in the lab, since observing only one organism removes it from its natural environment, which is a complex community of interacting organisms. Therefore, one cannot observe the organism's complete function or activity within lab conditions, since some gene will not be expressed, when there is no need to.

Def. operational taxonomic unit (=: OTU): A proxy for unknown species in the context of metagenomic analyses.

Ways to analyze microbial diversity:

Analysis of rRNA (16S rRNA in prokaryotes and 18S rRNA in eukaryotes – less does not yield unique results and more is often more difficult to analyze)

Rarefaction graph: Used to approximate the diversity of the found species in the community.
Mathematical form: rarefaction(#samples) = #found_species. Then plot #found_species versus #samples => the smaller the slope, the less new OTUs are identified.
This means, at the beginning, the slope will be quite steep, when #samples is low, but as #samples increases, there will be fewer and fewer #found_species.

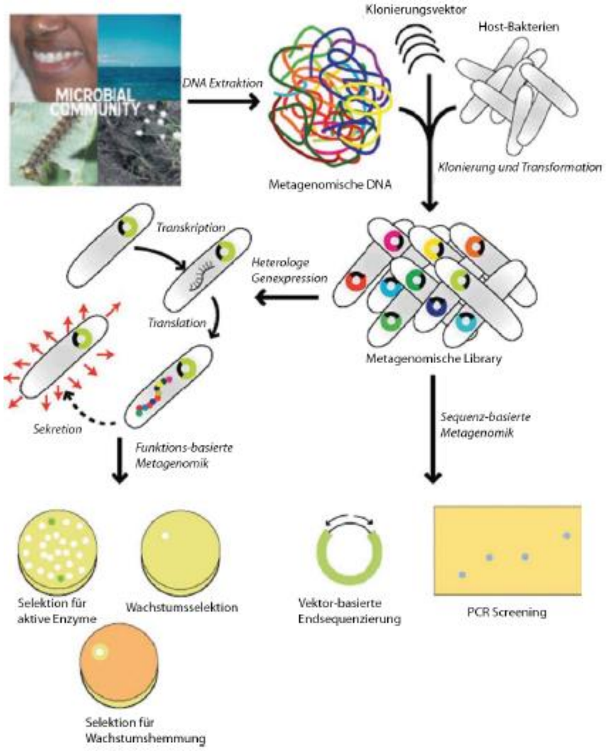
Metagenomic libraries

The DNA of a metagenome is often contaminated. Contamination is somewhat lower in water samples than in ground samples, since there are more DNases and humic acids. The purification process hinders the cloning of the metagenome or truncate the genetic information.
A gelelectropherosis separates the high molecular weight DNA fragments from humic acid, DNases and smaller fragments. Repeating extraction methods increases the rate of success of cloning and removes more humic acids and DNases (though some fraction of DNA will always be lost in each step). Another obstacle are spores or encapsulated bacteria that require very heavy lysis methods.

Types of cloning vectors:
Plasmids: around 15 kb
Cosmids and fosmids: around 40 kb
Artificial bacterial chromosomes: over 40 kb

Fosmids are present as a single copy in a cell (more stable than cosmids). 20 – 70 copies of cosmids are present in a cell.

Libraries with small inserts (up to 10 kb) require 3-20 times more clones than libraries with inserts in cosmids, fosmids or articifical bacterial chromosomes to cover the same diversity.



Def. electroporation: Making the cell membrane permeable temporary so that DNA can be introduced to a prokaryotic cell (transformation) or to a eukaryotic cell (transfection).

Alternatively, one can make use of phages to introduce metagenomic information to E.coli for example (transfection). Adding genes of organisms, that are distantly related to E.coli, proves to be problematic (try to find other suitable hosts for the expression of metagenomic data), because there are normally much fewer clones.

Sequence based metagenomics: Advantages: Gene sequence is output; genes can be found that are not expressed in bacterial hosts, for which there are no assays as libraries that have new functions potentially.

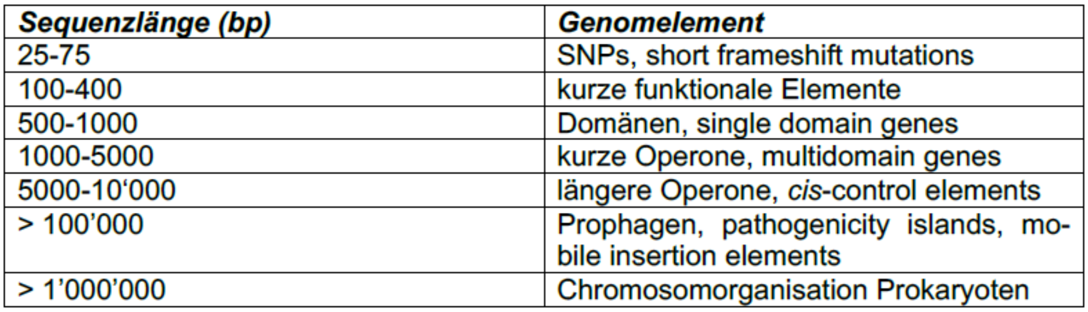
Functionally based metagenomics: Allows the identification of unknown biomolecules without prior knowledge of homologs etc. One identifies hosts (clones) that exhibit a particular metabolic activity and compares its metabolic activity to a metagenomic library.

Available sequencing technologies: Shotgun sequencing is used for a prokaryotic genome. Shotgun metagenomics provides information about the metabolic potential of the community and about the organisms present in the sample. It can to some extent cover underrepresented organisms in spite of its random nature.

High-through put sequencing: 454 pyrosequencing.
Also, SOLiD, MiSeq and HiSeq, ion torrent personal genome machine system etc. They generate smaller fragments but have higher reads than in Sanger sequencing. Another advantage of those technologies is that it is not necessary to prepare clones first.

Assembling

Assembling the DNA fragments yields information on ORFs, operons, operational transcriptional units and binding spots of transcription factors.



Sequenzlänge (bp)	***Genomelement***
25-75	SNPs, short frameshift mutations
100-400	kurze funktionale Elemente
500-1000	Domänen, single domain genes
1000-5000	kurze Operone, multidomain genes
5000-10‘000	längere Operone, *cis*-control elements
> 100’000	Prophagen, pathogenicity islands, mobile insertion elements
> 1’000’000	Chromosomorganisation Prokaryoten

Let L be read length, N number of reads, G the genome size, and C the coverage. Then,

$$C = \frac{L \times N}{G}$$

The percental part of the genome that is covered by the sequences is P_0:

$$P_0 = 1 - e^{-C} = 1 - e^{-(\frac{LN}{G})}$$

Also, $N = -\frac{\ln(1 - P_0)}{L} \times G$.

The metagenome size G_m of a sample with l species and n_j copies is:

$$G_m = \sum_{i=1}^{l} n_i G_i$$.

Also, one can write the following with p_i as the relative probability:

$$\hat{G}_m = p_1 G_1 + p_2 G_2 + \dots + p_l G_l \qquad \sum_{i=1}^{l} p_i = 1$$.

The p_i values and the diversity can be approximated with the proper gene markers (such as rDNA).

Def. binning: The mapping of sequencing data to OTUs.

Taxonomy dependent methods: Reads with sequences are compared to a reference data bank and are aligned. If a read has a too low similarity with the reference data bank, it is classified as unassigned.

Taxonomy dependent methods are classified as alignment-based, composition-based or as a hybrid method.
Alignment-based methods use BLAST to align individual reads of the sample with already known genomes. Finally, reads are classified through their alignments with different Hit-sequences from the database in taxonomic groups. Since those reads are often unknown, they are not assigned to the best BLAST hit, but to their **lowest common ancestor** (=: LCA).

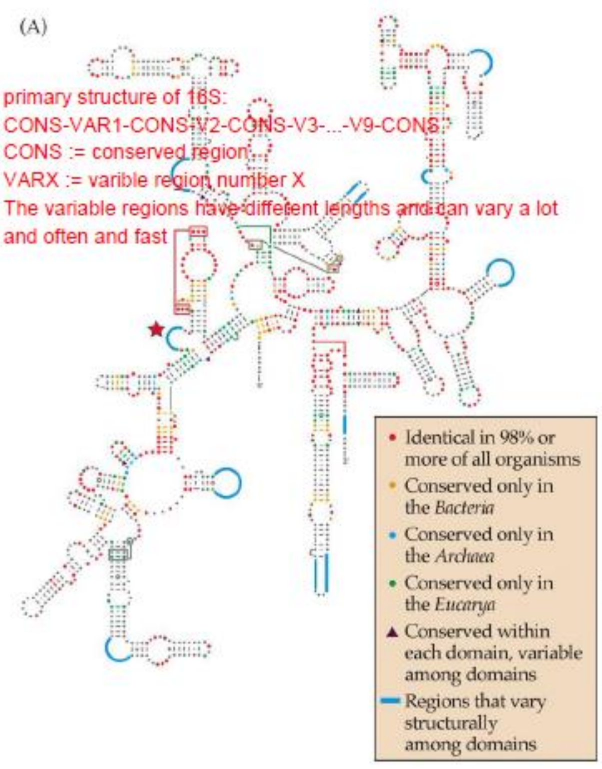
Composition based methods make use of GC-contents, codon usage, tetranucleotides (combination and relative abundance of four nucleotides, such as GGAG or GGAC etc.), oligonucleotide patterns to compare reads with sequences in databases. The final taxonomic classification is done by relative or absolute similarity. These methods are often quicker than alignment based methods and they require less computation power. Those reads have to be long enough though in order to make a sophisticated statement.

Taxonomy independent methods require no reference database and they only use the intrinsic information to cluster the reads in groups.

30.5.2017

Def. microbiome: The totality of all microorganisms in a community and their genomes.

Closely related bacteria have similar 16S rRNA sequences.


(A)
● Identical in 98% or more of all organisms
● Conserved only in the *Bacteria*
● Conserved only in the *Archaea*
● Conserved only in the *Eucarya*
▲ Conserved within each domain, variable among domains
— Regions that vary structurally among domains

primary structure of 16S:
CONS-VAR1-CONS-V2-CONS-V3-...-V9-CONS
CONS := conserved region
VARX := varible region number X
The variable regions have different lengths and can vary a lot and often and fast

MICROBIAL LIFE, Figure 17.3 (Part 1) © 2002 Sinauer Associates, Inc.

Disadvantageous of phenotypic screens:

Hits in screens are a lot more seldom than mathematically expected:

DNA can be from exotic bacteria that are only distantly related to E.coli. So, promoters and regulators are not recognized and codon usage can be different than in E.coli (rare codons). Sometimes, E.coli cannot fold the enzymes or post-translationally activate.

Solution: Choose other host or carry out DNA-screen than targeted expression.