Genetik, Genomik, Bioinformatik – FS18

v1

Gleb Ebert

13. August 2018

Diese Zusammenfassung soll den prüfungsrelevanten Stoff der Vorlesung Genetik, Genomik, Bioinformatik (Stand Frühjahrssemester 2018) in kompakter Form enthalten und basiert stark auf den Skripten des CAL. Ich kann leider weder Vollständigkeit noch die Abwesenheit von Fehlern garantieren. Für Fragen, Anregungen oder Verbesserungsvorschlägen kann ich unter `glebert@student.ethz.ch` erreicht werden. Die neuste Version kann stets unter `https://n.ethz.ch/~glebert/` gefunden werden.

Inhaltsverzeichnis

1 Genomik

Genomik ist die Anwendung von biologischen Forschungsansätzen und Labormethoden auf der Ebene des gesamten Genoms. Statt zum Beispiel die Funktion eines einzelnen Gens zu untersuchen, werden direkt alle Gene eines Organismus betrachtet. Die technischen Details der Methoden sind einem schnellen Wandel unterlegen. Daher werden sie oft an externe *core facilities* oder Firmen ausgelagert.
Eins der Hauptziele der Genetik ist es zu verstehen, wie der Genotyp den Phänotyp beeinflusst. Das Sequenzieren an sich ist heutzutage relativ einfach und konstengünstig. Die Speicherung, Verwaltung, Analyse und Interpretation stellen aber weit grössere Herausforderungen dar.

1.1 PCR-basierte Genotypisierung

PCR-basierte Genotypisierung wird verwendet, um Variation an einer bestimmten Stelle im Genom zu untersuchen (z.B. SNPs). Es können zum Beispiel zwei PCRs (*polymerase chain reaction*) mit leicht unterschiedlichen Primern durchgeführt werden. Die jeweils anderen Primer müssen aber identisch sein (Forward- oder Reverse-Primer). Primer, die am SNP eine andere Base enthalten, binden schwach und die PCR fällt entsprechend weniger stark aus.

1.1.1 TaqMan-Methode

Die Fluoreszenz-basierte TaqMan-Methode basiert auf einer PCR mit vier Primern. Zwei davon sind gewöhnlich, während die anderen beiden (A und B) mit je einem unterschiedlichen Fluorophor und einem Quencher-Molekül markiert sind. Durch die räumliche Nähe unterdrückt der Quencher die Floureszenz. Die markierten Primer binden direkt an den zu genotypisierten Locus. A bindet perfekt an Variante A und weisst einen Einzelbasenmismatch zu Variante B auf. Primer B ist entsprechend umgekehrt ausgelegt. Beim ersten Zyklus werden vorallem Produkte der einen Variante gebildet. Im darauf folgenden Zyklus bindet Primer A oder B and das Produkt und wird von der Polymerase zerschnitten. Der Quencher unterdrückt nun nicht mehr das Fluorophor und man kann die Fluoreszenz messen. Die relative Intensität der beiden Farben erlaubt Aussagen über den Genotyp des Individuums.
Mithilfe von real-time PCR-Geräten, kann der Verlauf der Fluoreszenz in Echtzeit mitverfolgt werden. Gegenüber herkömmlicher PCR bietet diese Methode den Vorteil von einer einzigen Reaktion und das Weglassen des Elektrophoreseschrittes, was beides äussere Einflüsse minimiert.

1.2 Microarray-basierte Genotypisierung

Moderne Microarray-basierte Methoden bestimmen den Genotyp an rund 1 Million Loci. Dabei werden parallel alle Loci auf einem sogenannten SNP-Chip genotypisiert.

Bsp. Illumina Infinium II
Rund 50bp lange, einzelsträngige DNA-Moleküle, sogenannte Oligos, werden auf einer Oberfläche in Clustern mit der selben Sequenz immobilisiert. Die Oligos sind jeweils komplementär zu über das Genom verteilten Loci. Wird die DNA-Probe nun zerkleinert, binden die Fragmente an ihre komplimentären Oligos, so dass man Cluster von doppelsträngigen DNA-Segmenten erhält. Das zu genotypisierte Nukleotid liegt direkt nach dem Nukleotid, welches an das 5'-Ende des Oligos bindet. Eine Polymerase baut nun an das 5'-Ende ein markiertes Nukleotid ein. Dieses kann anschliessend per optischen *read out* bestimmt werden. Der gesuchte SNP ist komplimentär dazu.

1.3 Gezielte DNA-Sequenzierung

Sucht man nach *de novo* Mutationen in proteinkodierenden Regionen, reicht es das Exom (ca. 1.5% des Genoms) zu sequenzieren um Zeit und Kosten zu sparen. Dazu werden nach der Zerstückelung der genomischen DNA Oligos mit zu proteinkodierenden Abschnitten komplementären Sequenzen beigegeben. Die Oligos sind mit Biotin markiert, welches die Bindung an Streptavidin-dekorierte Beads erlaubt. Nicht gebundene DNA wird ausgewaschen. Ein weiterer Denaturierungsschritt erlaubt das Sammeln der gesuchten DNA-Fragmente.

1.4 Genexpressionsstudien

Genexpressionsverändernde Mutationen rufen oft besonders starke Veränderungen des Phänotyp hervor. Mithilfe von Microarrays ist es heutzutage möglich, die Expression aller Gene eines Organismus in einem Versuch zu messen. Bei dieser Art von Microarray entsprechen die Oligos den zu untersuchenden mRNAs. Eine reverse Transkriptase synthetisiert die komplementären cDNAs, welche danach mit Fluoreszenzmarkern versehen werden. Die markierten cDNAs werden auf den Microarray gegeben. Anschliessend kann die relative Genxepression über die Intensität der Fluoreszenz gemessen werden. Die Bestimmung absoluter Expressionslevel ist aus technischen Gründen weniger zuverlässig.
Die sogenannte RNA-Seq Technologie löst Microarrays immer mehr ab. Dabei werden mit reversen Transkriptase wieder cDNAs aus mRNAs sequenziert. Nun werden aber die cDNAs direkt mit Hochdurchsatzsequenzierung analysiert. Die erhaltenen Sequenzen werden anschliessend per Alignment an das Genom angeglichen. Stark exprimierte Gene generieren viele *reads*.

1.5 Protein-DNA Interaktionen und Chromatinstruktur aus Sequenzierdaten

Sequenzierung kann auch für Fragestellungen verwendet werden, die nichts direkt mit der DNA-Sequenz zu tun haben.

1.5.1 Chromatin Immunoprecipitation and Sequencing

ChiP-Seq-Experimente untersuchen, an welchen Stellen ein bestimmtes Protein, z.B. ein Transkriptionsfaktor, mit dem Genom interagieren. Dazu gibt man das *crosslinking* Reagenz Formaldehyd einer lebenden Zelle bei. Es verursacht kovalente aber reversible chemische Bindungen zwischen benachbarten Proteinen und Nukleinsäuren. Anschliessend wir die Zelle aufgebrochen und die DNA zerstückelt. Beads mit Antikörpern binden nun das gesuchte Protein und der Rest wird ausgewaschen. Die *crosslinking* Reaktion wird rückgängig gemacht, die nun freien DNA-Fragmente sequenziert und per Alignment auf das Genom angeglichen.

1.5.2 Chromosome Conformation Capture

3C-Experimente, z.B. Hi-C, verwenden *crosslinks* zwischen benachbarten DNA-Abschnitten. Durch unspezifische Re-

striktionsenzyme wird die DNA in Fragmente zerteilt. Einzelstrangüberhänge werden mit Biotin-markierten Nukleotiden aufgefüllt und die so enstandenen *blunt ends* von einer Ligase verknüpft. Vorher weit auseinander liegende DNA-Abschnitte liegen nun in einem linearen DNA-Molekül nebeneinander. Diese Abschnitte waren vorher in der räumlichen Chromatinstruktur nah beieinander. Nach erneutem Zerstückeln werden die Fragmente über Streptavidin-Beads aufgereinigt und sequenziert. Die beiden Enden der Hybridsequenzen sind dann an verschiedene Stellen des Genoms angleichbar, woraus geschlossen werden kann, dass diese Abschnitte im Zellkern nah beieinander liegen.

1.6 DNA-Synthese

Die Synthese von Oligonukleotiden ist den Sequenziermethoden der 2. und 3. Generation nicht unähnlich. Die Oligos sind auf einer Oberfläche mit bestimmten xy-Koordinaten fixiert. Es gibt verschiedene Techniken, um das gewollte Nukleotid an ein bestimmtes Oligo anzuhängen.
Die Firma Agilent z.B. verwendet die ursprünglich für Tintenstrahldrucker entwickelte Inkjet Technologie. Dabei wird auf jede xy-Position ein winziges Reagenztröpchen gegeben. Ein anderer Ansatz der Firmen Affymetrix, Nimblegen und CustomArray ist es, den Chip nacheinander mit jedem der vier Nukleotide zu fluten und aber nur an den gewollten Positionen die Reaktion erlauben. Die Aktivierung durch photochemische Entfernung von Blockermolekülen geschieht über *micromirror arrays*, die genaue Steuerung von Lichtstrahlen ermöglichen. Alternativ erlauben Chips mit ansteuerbaren Elektroden, welche den pH an der Oberfläche verändern können, die elektrochemische Entfernung der Blocker.
Derzeit ist die Länge der Oligos auf 75 bis 200 Basen beschränkt und die Fehlerrate liegt bei rund 0.1%.

2 Bioinformatik

Die Bioinformatik hat sich durch den rasanten Zuwachs der Datenmengen aus verschiedenen Omics-Techniken zu einem immer wichtigeren Bestandteil der biologischen Forschung entwickelt. Es ist jedem zu empfehlen zumindest eine Programmiersprache wie Python, Perl, Matlab oder R zu lernen.

2.1 Modelle

> *„all models are wrong, but some are useful“*
> — George P. Box

Modelle sind unterschiedlich gute Versuche, biologische Realitäten in mehr oder weniger vereinfachter Form darzustellen. Bioinformatische Analysemethoden verwenden Algorithmen, die wiederum auf Modellen basieren. Schwierigkeiten entstehen dabei meist aufgrund unpassenderer Anwendung von Algorithmen.

2.2 Datensätze, Datenbanken und Tools

2.2.1 Genom-Projekte

Vorhaben wie das Human Genome Project wollen kontinuierliche Sequenzen aller Chromosomen eines Organismus in einem sogenannten *build* oder *assembly* zusammensetzen. Die grösste Herausforderung sind dabei mehrfach auftretende, homologe Sequenzen. Zum Beispiel das menschliche Genom weist bis heute Bereiche auf, in denen die genaue Abfolge der Sequenzen unbekannt ist. Aus diesem Grund spricht man oft von Genom-Entwürfen oder *draft assemblies*.

2.2.2 1000 Genomes

1000 Genomes (`http://www.internationalgenome.org/`) dokumentiert die Variationen zwischen der Vielzahl an menschlichen Genomen.

2.2.3 dbSNP

dbSNP sammelt alle bekannten Einzelbasenvariationen (SNPs, Insertions, Deletions) des menschlichen Genoms.

2.2.4 KEGG

Die Kyoto Encyclopedia of Genes and Genomes ist vorallem für Informationen zu Stoffwechselprodukten und -wegen bekannt. `https://www.genome.jp/kegg/`

2.2.5 GenBank

GenBank entählt Sequenzen mit einer Gesamtlänge von fast einer Trillion Basenpaaren. Sie sind alle mit Informationen über Spezies, Ursprung, potentielle Gene, etc. versehen und haben eine individuelle Identifikationsnummer (*accession number*). Der RefSeq-Datensatz ist nicht redundant, enthält jedes Chromosomensegment, jede RNA und jedes Protein also nur ein Mal. Er ist besonders nützlich als Referenz.

2.2.6 ENCODE

Die Encyclopedia of DNA Elements hat als Ziel die komplette Annotation des menschlichen Genoms mithilfe verschiedener High-Throughput-Technologien.
`https://genome.ucsc.edu/ENCODE/`

2.2.7 Bioinformatik-Tools im Web

Das EBI (European Bioinformatics Institute) und das NCBI (National Center for Bioinformatics) bieten viele verschiedene Webtools für immer komplexere Aufgaben.
`https://www.ebi.ac.uk/services/all`

2.2.8 Pubmed

Pubmed ist eine Suchmaschine für biologisch-medizinische Literatur.

2.2.9 Genome-Browser

Genome-Browser sind interaktive Websiten, die Informationen aus Datenbanken sammeln und übersichtlich darstellen. Einer der meistbenutzen ist derjenige der University of California Santa Cruz. `https://genome.ucsc.edu/`

2.2.10 APIs

Sogenannte *application programming interfaces* sind Schnittstellen, die es direkt über Programmiersprachen erlauben, auf Datenbanken zuzugreifen. Dabei kann der Syntax einfach angepasst werden, so dass man sich nicht in verschiedene Datenbanken einarbeiten muss. Das Buch Entrez Programming Utilities Help (`https://www.ncbi.nlm.nih.gov/books/NBK25501/`) bietet weiterführende Informationen.

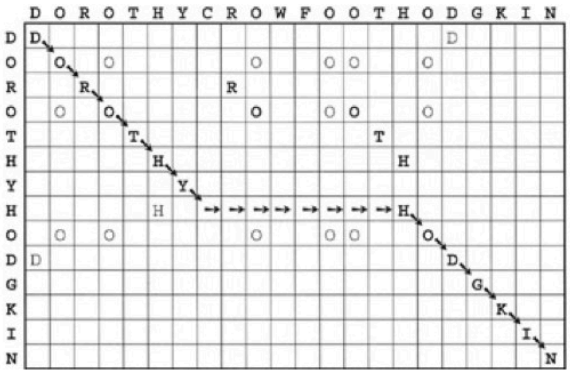
2.3 Sequenzalignment

Ein Alignment ist die Eins-zu-eins-Zuordnung von Nukleotiden oder Aminosäuren zwischen verschiedenen Sequenzen. Dabei gibts es paarweise und multiple (drei oder mehr Sequenzen) Alignments. Gute Alignments entstehen durch gute Balance folgender Qualitätskriterien:

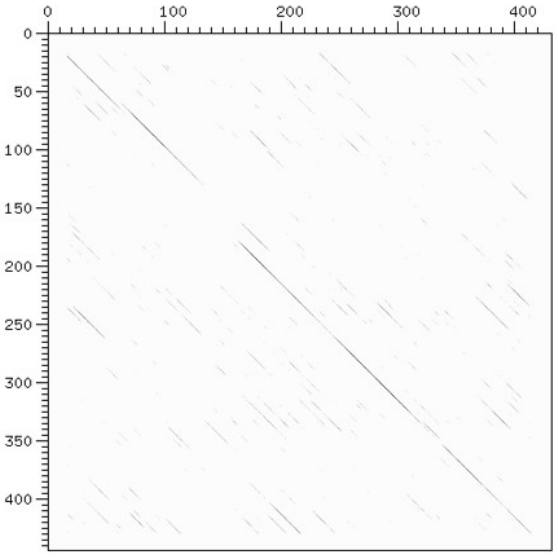
- viele *matches* bzw. wenige *mismatches*
- wenige *gaps* (Lücken bzw. Leerstellen)
- kurze *gaps*

2.3.1 Dot-Plots

In einem Dot-Plot wird die erste Sequenz auf die x- und die zweite Sequenz auf die y-Achse aufgetragen. Stimmen die Sequenzen in einer Position überein, wird diese mit einem Punkt (*dot*) markiert. Jeder Pfad, der von der oberen linken zur unteren rechten Ecke führt und dabei nur Schritte in positive Richtung der Achsen macht, entspricht einem Alignment.




```
DOROTHYCROWFOOTHODGKIN
|||||||         |||||||
DOROTHY--------HODGKIN
```



2.3.2 Bestes Alignment finden

Eine Bewertungsfunktion berechnet für jedes Alignment anhand der Anzahl *mismatches* und *gaps* einen Zahlenwert, der die Qualität wiederspiegelt. Die Bewertungsfunktion hängt dabei von der Fragestellung ab. Anschliessend wird über einen Suchalgorithmus das beste Alignment bestimmt. Mit sogenanntem *dynamic programming* kann mathematisch beweisbar das beste Alignment gefunden werden.

2.3.3 Globale und lokale Alignments

Werden zwei gleichlange und relativ kurze Sequenzen miteinander verglichen, spricht man von globalem Alignment. Der berühmteste *dynamic programming* Algorithmus dafür ist der Needleman-Wunsch-Algorithmus.
Sind nur Teile der beiden Sequenzen ähnlich, werden lediglich diese einem lokalem Alignment unterworfen. Hier ist der Smith-Waterman-Algorithmus ein Beispiel für *dynamic programming*.

2.3.4 Heuristische Algorithmen

Bei grossen Alignmentproblemen sind *dynamic programming* Algorithmen zu langsam. Deswegen verwendet man heuristische Algorithmen, welche Annahmen treffen und Abkürzungen nehmen. Unter den richtigen Rahmenbedingungen kann ein solcher Algorithmus sehr gute Lösungen extrem schnell finden.

Bsp. BLAST Algorithmus:
BLAST (*basic local alignment search tool*) fängt damit an, dass *low complexity* Regionen rausgefiltert werden. Der Rest wird in Wörter (11 Nukleotide lang bei DNA) zerlegt. Für jedes Wort wird die Wahrscheinlichkeit berechnet, dass es durch Zufall auftritt. Anschliessend wird in der zu durchsuchenden Datenbank nach exakten Übereinstimmungen mit den weniger zufälligen Wörtern gesucht. Bei jeder der gefundenen Übereinstimmungen wird in beide Richtungen der Datenbanksequenz nach weiteren Übereinstimmungen gesucht. *Mismatches* sind dabei erlaubt. Eine Bewertungsfunktion entscheided jedes Mal, ob eine solche Verlängerung das gesamte Alignment verbessert oder nicht. Zum Schluss werden alle lokalen Alignments verknüpft, wenn diese nah genug beieinander liegen. Hier sind dann auch *gaps* erlaubt. Diese Methode ist vorallem aus zwei Gründen besonders effizient. Durch Filtern der *low complexity* Regionen wird der *search space* verkleinert. Ausserdem sind exakte Übereinstimmungen einfacher zu finden. Der geschwindigkeitsbestimmende Schritt ist beim BLAST Algorithmus die Wortlänge.

2.3.5 Alignment von Next-Gen-Sequencing-Daten

Wenn *next generation sequencing* Technologien für *whole genome sequencing* verwendet werden, erhält man eine Unzahl an rund 100bp langen Fragmenten. Um Zeit zu sparen, wird für das Alignment dabei ein Index, z.B. der bei langen DNA-Sequenzen besonders effiziente Burrows-Wheeler-Index, verwendet. Dieser ist vereinfacht gesagt ein Register mit sortierten Einträgen, in denen man viel schneller den gesuchten findet. Der Index selber ist relativ aufwändig zu generieren, kann aber vielfach benutzt werden. Zum Index gibt es jeweils eine entsprechende Methode, z.B. das Burrows-Wheeler-Alignment.

2.4 Multiples Sequenzalignment (MSA)

Der parallel Vergleich mehrerer Sequenzen ist besonders praktisch, um funktional wichtige Sequenzteile zu finden. Tauchen in diesen nämlich *loss-of-function*-Mutationen auf, kann sich der Genotyp meist nicht in der Population halten und der Teil bleibt weitgehend konserviert. Je grösser der evolutionäre Abstand zwischen zwei homologen Sequenzen ist, desto schwieriger ist diese Homologie zu erkennen. Bei Proteinsequenzen ist dies aber einfacher, da man bei 20 Aminosäuren einen höheren Informationsgehalt gegenüber 4 Basen hat.

2.4.1 Progressive Alignment-Algorithmen

Da formal beweisbare Algorithmen für MSA rechnerisch zu aufwendig sind, greift man auf sogenanntes progressives Alignment, z.B. CLUSTALW, zurück. Als erstes werden die beiden sich am meisten ähnelnden Sequenzen bestimmt. Anschliessend werden weitere Sequenzen an das erste Paar angeglichen. Dabei ist das grösste Problem, dass Fehler im ersten Alignment später nicht mehr ausgebessert werden können.

2.4.2 Iterative Alignment-Algorithmen

Iterative Alignment-Algorithmen umgehen das Problem der progressiven, indem nach jedem Alignment das erste Paar überprüft und wenn notwendig angepasst wird. Eine besonders erfolgreiche Implementierung heisst MUSCLE und ist online verfügbar: `https://www.ebi.ac.uk/Tools/msa/muscle/`.

2.4.3 Hidden Markov Model (HMM) Algorithmen

HMM Algorithmen basieren auf einer Klasse von statistischen Modellen und sind ohne weitreichende Mathematikkentnisse nicht einfach verständlich. Sie liefern allerdings heutzutage die besten Alignments. Weit verbreitete Implementationen sind HMMER oder ClustalOmega (`https://www.ebi.ac.uk/Tools/msa/clustalo/`).

2.4.4 Profil-basierte Algorithmen

Profil-basierte Algorithmen sind besonders gut darin, kurze aber stark konservierte Motive zu finden. Sie können selbst stark divergierte Sequenzen einer bestimmten Funktion zuordnen.

2.5 Phylogenetische Bäume

Ein phylogenetischer Baum zeigt grafisch evolutionäre und funktionale Verwandtschaften zwischen mehreren Sequenzen. Dabei ist zu beachten, dass aufgrund von Prozessen wie Genkonversation, -duplikation, -transfer oder funktionaler Selektion molekulare Bäume nicht direkt aus phylogenetischen Bäumen abgelesen werden können.
Die Entfernung der Sequenzen vom letzten gemeinsamen Vorfahren kann dabei gleich lang gehalten werden um evolutionäre Prozesse darzustellen. Alternativ kann die Länge der Äste proportional zur Anzahl Mutationen zwischen zwei Punkten sein.

2.5.1 Maximum-Likelihood-basierte Algorithmen

Diese Algorithmen bilden alle möglichen Bäume und berechnen dann jeweils ihre Wahrscheinlichkeit. Bei einer grossen Anzahl Sequenzen ist dieser Ansatz allerdings sehr rechenaufwändig.

2.5.2 Parsimony-basierte Algorithmen

Bei dieser Methode wird der Baum mit den wenigsten evolutionären Schritten, die benötigt werden, um von der Ursprungssequenz zu allen beobachteten Sequenzen zu gelangen, ausgewählt. Solche Algorithmen sind besonders gut geeignet bei eng miteinander verwandten Sequenzen.

2.5.3 Entfernungs-basierte Bäumen

Sogenannte *distance based methods* liefern einen guten Kompromiss aus Zuverlässigkeit und Geschwindigkeit. Dabei berechnet der Smith-Waterman-Algorithmus einen *score* für jede mögliche Entfernung zwischen je zwei Sequenzen. Bei mehr als drei Sequenzen sind die Verhältnisse oft nicht exakt zweidimensional darstellbar. Algorithmen wie UPGMA (*unweighted pair-group method with arithmetic mean*) treffen Kompromisse unter der Annahme, das die Mutationsrate konstant ist.

2.5.4 Outgroup

Erfahrung zeigt, dass eine etwas weiter entfernte Sequenz, die sogenannte *outgroup*, das Resultat der Algorithmen wesentlich verbessert.

2.6 Clustering

Clutstering-Algorithmen ordnen gleiches zu gleichem und sind sehr hilfreich bei grossen Datenmengen. Die Resultate reflektieren dabei das gewählte Modell. Ansätze sind Agglomeration, die *bottom-up* ist, viel Variation aufweist, häufig für Genexpressionsdaten verwendet wird und aber zeitintensiv für grosse Datensätze ist (n^2) als auch Partition, der *top-down* ist, den Datensatz in ein vorgegebene Anzahl Cluster zerlegt und zeiteffizient für grosse Datensätze ist (n). Oft liefern Hybridlösugen bessere Resultate als eine der Methoden alleine.

2.6.1 Agglomerative hierarchical clustering

1) vergleiche alle Datenvektoren miteinander
2) verknüpfe diejenigen die einander am ähnlichsten sind
3) vergleiche die entstandenen Cluster und Vektoren miteinander
4) wiederhole **2)** und **3)**

Zwei beliebte Metriken für Ähnlichkeit sind die euklidische Distanz (Länge des Differenzvektors) und der Persons Korrelationsfaktor (*cos* des Winkels zwischen den Vektoren). Agglomeratives Clustering läuft nur in eine Richtung und der Nutzer muss viele Parameter wählen.

2.6.2 k-means Clustering

1) für jedes der k Cluster wähle per Zufall einen Clustervektor
2) ordne jeden Datenvektor einem Cluster zu, basierend auf Ähnlichkeit von Daten- und Clustervektor
3) Berechne aus den Datenvektoren im Cluster einen neuen Clustervektor (Mittelpunkt)
4) wiederhole **2)** und **3)**

3 Bakterielle Genetik

Einige Bakterienstämme sind besonders einfach im Labor zu handhaben, manipulieren und vermehren und sind somit gut als Modellsystem geeignet. Viele wichtige Erkentnisse der Biologie wurden mithilfe von Bakterien erarbeitet. Darunter sind viele fundamentale molekulare und zelluläre Prozesse, Mechanismen der Translation, molekulare Maschinen der DNA-Replikation, Chaperone und Mechanismen der Mutagenese und DNA-Reparatur.
Bakterien sind weitaus komplexer als man auf den ersten Blick vermuten könnte. So ist z.B. die Verteilung vieler Enzyme innerhalb der Zelle streng kontrolliert und die DNA des Genoms präzise strukturiert. Ausserdem spielen Bakterien eine zentrale Rolle in vielen Ökosystemen und der menschlichen Gesundheit.

3.1 Das bakterielle Genom

Bakterielle Genome sind relativ klein. Die Länge reicht von nur rund 160kb (0.16Mb) bis 10Mb. Weil Introns und repetitive Sequenzen kaum zu finden sind, ist die Gendichte aber rund 100-mal höher als beim Menschen (im Durchschnitt 1 Gen alle 1.1 kb). Die meisten Sequenzen kodieren für Proteine, rRNA und tRNA. Es werden aber immer mehr für kurze Peptide und regulatorische RNAs kodierende Sequenzen gefunden.

3.2 Plasmide

Plasmide sind zirkuläre doppelsträngige DNA-Moleküle mit einigen tausend bis hunderttausend Basenpaaren. Vorallem Gene, die einem Bakterien unter spezifischen Bedingungen einen Vorteil verschaffen, sind darauf vorhanden. Plasmide können durch Konjugation an andere Bakterienzellen – auch von anderen Arten oder sogar an Eukaryoten – übertragen werden. Aufgrund all dieser Eigenschaften sind Plasmide hervorragende Werkzeuge der Mikro- und Molekularbiologie.

3.3 Vorteile der bakteriellen Genetik

- Bestimmte Bakterienarten wie *E. coli* lassen sich sehr einfach genetisch manipulieren.
- Das bakterielle Genom ist haploid und enthält von jedem Gen also nur eine Kopie bzw. ein Allel. Zellen mit bestimmten Mutationen können so besonders einfach identifiziert werden, da der Effekt sofort sichtbar wird.
- Bakterien haben eine kurze Generationsdauer (20 Minuten bei *E. coli*).
- Bakterien vermehren sich durch Zellteilung, also asexuell. Alle Nachfahren sind demnach genetisch identisch und werden Klone genannt.
- Bakterien wachsen in Kolonien zu Tausenden, Millionen oder sogar Milliarden auf Agarplatten. Ausserdem stammen alle Zellen innerhalb einer Kolonie von einer einzigen Vorgängerzelle.
- Durch Selektion können Mutationen oder Stämme isoliert werden. Dazu wählt man spezifische Wachstumsbedingungen, unter denen nur die gesuchten Zellen wachsen können.

3.4 Synthetische Genomik

Inzwischen ist es möglich gesamte bakterielle Genome sythetisch herzustellen. In Zukunft könnte es also möglich sein, Bakterien für bestimmte Anwendungen masszuschneidern. Allerdings stellt sich dann die Frage nach der Ethik und ob genug Sicherheitsmassnahmen vorgenommen werden können.

3.5 Variabilität in Bakterien

3.5.1 Physiologische Variation

Physiologische Variationen treten aufgrund der leicht unterschiedlichen Umgebung, einer leicht unterschiedlichen Wachstumsgeschichte und der momentanen Zellzyklusphase jeder einzelnen Bakterienzelle auf. Diese Variation wird nicht vererbt.

3.5.2 Genetische Variation

In eukaryotischen Zellen ist die Hauptquelle der genetischen Variabilität die Meiose. Da sich Bakterien aber asexuell vermehren, kommt die genetische Variation von Mutationen, Rekombination und Austausch genetischer Informationen. Populationen von Bakterien sind niemals wirklich zu 100% genetisch homogen.

3.6 Mutationsraten

Die Molekular Mutationsrate ist die Häufigkeit von Mutationen pro DNA-Replikationszyklus. Sie wird in Anzahl Mutationen pro Anzahl replizierte Basenpaare angegeben. Die phänotypische Mutationsrate beschreibt die Wahrscheinlichkeit, dass in einem bestimmten Zeitintervall eine Mutation zu einem bestimmten Phänotyp auftritt. Sie ist besonders hoch, wenn mehrere molekulare Mutationen den selben Phänotyp auslösen (Bsp: siehe Histidin-Auxotrophie und Streptomycin-Resistenz).

Mutationstyp	$\frac{\text{Mutationsrate}}{\text{Nukleotid} * \text{Generation}}$
Basenpaarsubstitution	200×10^{-12}
stille Mutationen	50×10^{-12}
Missense-Mutationen	150×10^{-12}
Nonsense-Mutation	6×10^{-12}
Indels (< 4 bp)	20×10^{-12}

3.7 Mutationsarten

- Basenpaarsubstitution: Ein Basenpaar wird durch ein anderes ersetzt.
- Stille Mutation: Mutation in einer kodierenden Region, die aber das Protein nicht verändert.
- Neutrale Mutation: Alle Mutationen, die keinen phänotypischen Effekt haben.
- Missense-Mutation: Im Protein wird eine Aminosäure durch eine andere ersetzt. Diese Substitution muss aber keinen Effekt haben.
- Nonsense-Mutation: Ein Aminosäuren-kodierendes Codon wird in ein Stop-Codon (UAA, UAG, UGA) umgewandelt.
- Frameshift-Mutation: Basenpaare werden zu einem ORF hinzugefügt oder entfernt, nicht aber in Vielfachen von 3. Die Proteinfunktion kann nicht bis stark beeinflusst werden.

3.8 Klassische genetische Analyse in Bakterien

3.8.1 Isolierung von Mutanten

Isolierung ist der Prozess eine Mutante in einem Pool von nicht mutierten Organismen zu finden. Man muss den Phänotyp der Mutation kennen, um geeingete Selektionsbedingungen anwenden zu können.

3.8.2 Screenen nach Mutanten ohne Selektion

Da es oft keine Selektionsbedingungen für die Mutation gibt, werden Bedingungen angewendet, unter denen der Wildtyp aber nicht die Mutante wächst. Auf diese Art findet man Kolonien, welche die gesuchte Mutation aufweisen.

3.8.3 Selektion vs. Screening

> *„Eine Selektion ist besser als tausend Screens."*
> — David Botstein

Eine gute Selektion kann Mutanten mit einer Wahrscheinlichkeit kleiner als 10^{-10} finden. Screens haben eine kleinere *power of resolution* mit auffindbaren Wahrscheinlichkeiten von 10^{-2} bis 10^{-4}.

3.8.4 Zuordnung von Mutanten

Mithilfe von Referenzgenomen kann relativ einfach der genaue Ort der Mutation gefunden werden. Ausserdem können verwandte Gene gesucht werden, mit denen auf eine mögliche Funktion des mutierten Gens zu schliessen möglich ist.

4 Hefe-Genetik

4.1 Hefe als Modellorganismus

Sowohl Knospungshefen als auch Spalthefen sind gute Modellsysteme für eukaryotische Zellen, da die Zellstruktur und der Zellzyklus ähnlich sind. Des weiteren sind sie ein gutes Mittel um eukaryotische Gene zu entdecken und kategorisieren. Hefen finden ausserdem aufgrund ihrer Fähigkeit Zucker in Abwesenheit von Sauerstoff in Ethanol und CO_2 zu gähren und ihrer Toleranz gegenüber sauren und alkoholischen Umgebungen Anwendung in der Lebensmittelindustrie bei der Herstellung von Brot, Bier oder Wein.

4.2 Saccharomyces cerevisiae

Die Knospungshefe *Saccharomyces cerevisiae*, auch Back-, Bäcker- oder Bierhefe genannt, war an vielen gentechnologischen Meilensteinen beteiligt. Erste...

- Transformation einer eukaryotischen Zelle mittels Plasmiden
- Konstruktion genauer Gen-Knockouts für Eukaryoten
- komplette Sequenzierung eines Eukaryoten-Genoms

S. cerevisiae hat ein haploides Genom von 12 Millionen Basenpaaren (12Mb) mit 16 Chromosomen, die rund 6'200 Gene (davon etwa 5'800 Protein-kodierende) enthalten. 76% aller Gene sind sowohl charakterisiert als auch verifiziert, 12% nur charakterisiert und die restlichen 12% haben noch keine bekannte Funktion. Die hohe Gendichte ist durch wenig Introns und niedrige Redundanz zu erklären. Zusätzlich hat die Hefe ein mitochondriales Genom als auch das 2μ-Plasmid.

4.2.1 Lebens- und Zellzyklus

S. cerevisiae kann sowohl in Flüssigkultur als auch auf festem Medium gezüchtet werden und weist dabei eine Generationszeit von lediglich 90 Minuten auf. Der Lebenszyklus weist Wachstumsphasen sowohl im haploiden als auch diploiden Zustand auf. Konjugation ist die Verschmelzung zweier haploider Zellen zu einer diploiden und kann während Experimenten ausgelöst werden. Einzelne Zellen können wiederum isoliert werden, um isogene Zuchtlinien zu erhalten.

S. cerevisiae vermehrt sich vorallem asexuell durch Knospung. Die Mutterzelle bildet dabei einen Auswuchs mit einem eigenen Zellkern. Im sexuellen Zyklus kann wieder über Knospenbildung eine Tochterzelle oder durch Meiose und zwei Zellteilungen vier haploide Ascosporen gebildet werden. Diese vier Tochterzellen, auch Tetrade genannt, werden zunächst noch vom Ascus (eine Hülle) zusammengehalten und von der Aussenwelt geschützt. Die Sporen an sich sind aber auch schon um einiges resistenter gegenüber ungünstigen Umwelteinflüssen. Jede dieser Sporen hat ausserdem einen vom *MAT*-Locus bestimmten Paarungstyp (entweder a oder α). Die Tetrade ist mithilfe eines mit einem Micromanipulator ausgestattetem Mikroskop trennbar.

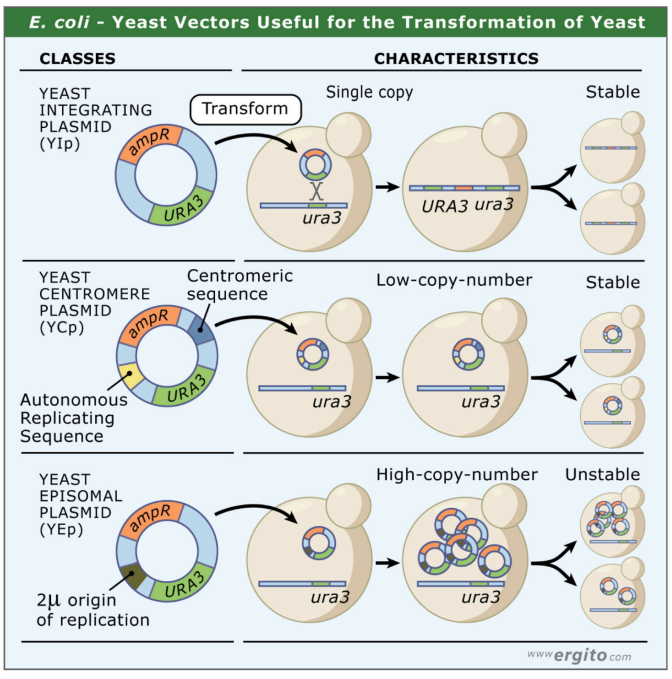
Die Morphologie der Zelle ändert sich im Laufe des Zellzyklus. In der G1-Phase ist sie rund und knospenlos. Sobald sich eine kleine Knospe gebildet hat, befindet sie sich in der S-Phase. In der G2-Phase wächst die Knospe und der Zellkern wandert in ihre Nähe. Zum Schluss läuft die Mitose ab.

4.3 Schizosaccharomyces pombe

Die Spalthefe *Schizosaccharomyces pombe* ist evolutionär sehr weit von *S. cerevisiae* entfernt: Ihr letzter gemeinsamer Vorfahre liegt rund 1'000 Millionen Jahre zurück. *S. pombe* ist ein wichtiges Modellsystem für Zellzyklus und Zellwachstum. Der stäbchenförmige Eukaryot verfügt über eine Generationszeit von 2 bis 4 Stunden und vermehrt sich durch klassische Zellteilung. Im Gegensatz zu *S. cerevisiae* hat *S. pombe* nur 3 Chromosomen, obwohl die Länge des Genoms mit rund 12Mb ähnlich ist. Nur ca. 4'800 Proteinkodierende Gene sind bei *S. pombe* im Gegensatz zu den 5'800 von *S. cerevisiae* zu finden. Die Gene sind bei den beiden Arten auch in einer unterschiedlichen Reihenfolge angeordnet, es besteht als keine Syntenie. Dies ist auf mehrere Genduplikationen gefolgt von Genverlust im Laufe der Evolution bei *S. cerevisiae* zurückzuführen. Duplizierte Gene konnten sich anschliessend in verschiedene Richtungen entwickeln und bildeten so Paraloge (verwandte Gene mit unterschiedlichen Funktionen). Ein weiterer Unterschied ist bei den Introns zu finden: Etwa 40% der *S. pombe* Gene haben welche, aber nur ca. 5% bei *S. cerevisiae*.

4.4 Genetische Screens vs. Selektion

Die meisten Konzepte sind analog zu denen in der bakteriellen Genetik (Kapitel 3.8). Des weiteren ist zu Screens anzumerken, dass durch das Untersuchen jeder Kolonie ein breiteres Spektrum an Mutanten mit verschiedenen Phänotypen identifiziert werden kann. Manche Arten von Mutanten können in Selektionen nämlich übersehen werden (wenn z.B. die Mutante nicht kompetitiv gegenüber dem Wildtyp ist). *S. cerevisiae* ist dank des kleinen Genoms und der kurzen Generationszeit gut zur Identifikation von Genen durch zufällige Mutationen geeignet.
Um Mutationen zu identifizieren kann entweder das Genom sequenziert und mit einem Referenzgenom verglichen werden, oder man führt eine Komplementation aus. Dabei probiert man Plasmide aus einer Plasmid-Bibliothek aus, bis man welche findet, die den Phänotyp komplementieren, d.h. das mutierte Allel retten.



4.5 Mutagenesemethoden

Mutagenese kann die Mutationsrate pro Gen um bis das 100-fache erhöhen. Sie bleibt jedoch ein zufälliger Prozess.

- Chemische Mutagenese durch Ethylmethansulfonat (EMS) verursacht Punktmutationen durch Methylierung der Basen, was zu fehlerhaftem Einbau während der Replikation führt. Sie ist besonders gut für temperaturempfindliche Mutationen geeignet (siehe Kapitel 4.6).
- Physische Mutagenese wird durch Bestrahlung mit UV-Licht verursacht und bewirk verschiedene Mutationen wie Transitionen oder Transversionen.

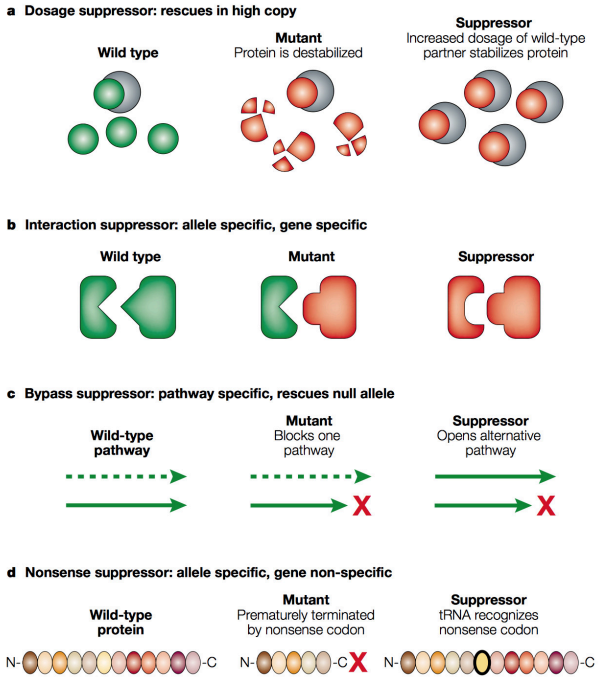
Findet man immer und immer wieder die selben Mutationen, geht man davon aus, dass man praktisch alle gefunden hat und der Screen somit gesättigt (*saturated*) ist.
Die yeast knoch-out collection (YKO) umfasst über 21'000 mutante Hefestämme. Sie enthält genaue Start-zu-Stop-Deletionen von rund 6'000 ORFs und erreicht so eine Abdeckung von etwa 96%.

4.6 Temperatursensitive Mutationen

Da *Loss of function*-Mutationen in essentiellen Genen tödlich sind, existieren keine Zuchtlinien. Deswegen wird nach schwächeren Varianten gesucht, die im besten Fall zusätzlich Temperatursensitivität mit sich bringen. Hitzeempfindliche (*thermosensitive, ts*) Proteine werden bei höheren Temperaturen inaktiv (manchmal umgekehrt bei tieferen). Kälteempfindliche (*coldsensitive, cs*) Mutationen beeinflussen oft Protein-Protein-Interaktionen bei tieferen Temperaturen.

4.7 Synthetische Letalität und Suppressionsanalyse

Bei der Suppressionsanalyse wird der Phänotyp einer Mutation durch eine zweite Mutation oder erhöhte Expression eines anderen Gens gerettet. Ihr Gegenteil ist synthetische Lethalität. Dabei wird der Phänotyp einer Mutation so von einer zweiten Mutation oder der Überexpression eines anderen Gens verstärkt, dass er tödlich wird.



4.8 Temperatursensitive Zellzyklus-Mutanten

Bsp. Genetischer Screen vom L.H. Hartwell in *S. cerevisiae*
Die Forscher suchten zunächst nach Mutanten, bei denen die Vermehrung bei 23°C noch funktionierte, aber bei 36°C zum erliegen kam. Durch Beobachtung wurde festgestellt, dass alle Zellen die selbe Morphologie aufwiesen, sprich sich in der selben Zellzyklusphase aufhielten, kurz nachdem die Temperatur erhöht wurde. Diese Mutanten wurden *cell cycle division (cdc)*-Mutanten genannt.

Paul Nurse konnte zeigen, dass eine *cdc2*-Mutante von *S. pombe* mit einem menschlichen *cdc2*-Äquivalent gerettet werden konnte.

5 Genetische Studien in Drosophila

5.1 Drosophila als Modellorganismus

Drosophila melanogaster ist ein einfacher, mehrzelliger Modellorganismus, der zur Identifizierung von an Zelldiferenzierung und Wachstum beteiligten Genen, als auch zur Untersuchung der Embryonalentwicklung benutzt wird. Männchen und Weibchen sind von Auge unterscheidbar. Das Genom besteht aus vier Chromosomenpaaren, drei davon Autosomen und ein Paar Geschlechtschromosomen. Die geringe Anzahl macht es einfacher rauszufinden, auf welchem Chromosom ein Gen bzw. eine Mutation liegt. 50% der etwas 15'000 Gene sind homolog zum Menschen, obwohl sich die Entwicklungslinien vor über 700 Millionen Jahren getrennt haben. Die Generationszeit beträgt lediglich 10 Tage vom Ei zur Fliege und die Haltung ist einfach und kostengünstig. Die Entwicklung zur adulten Fliege läuft über drei Larverstadien und ein Puppenstadium ab.

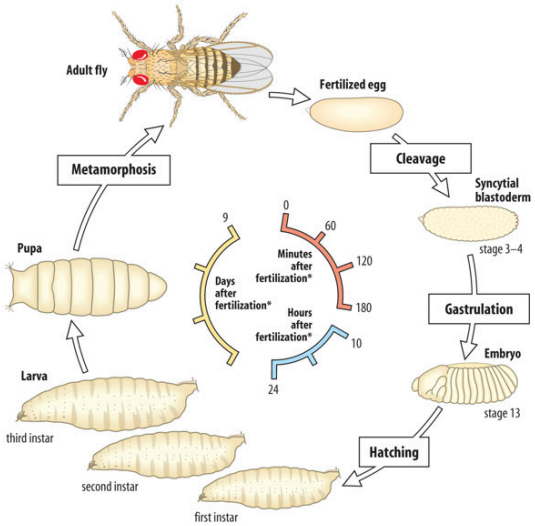
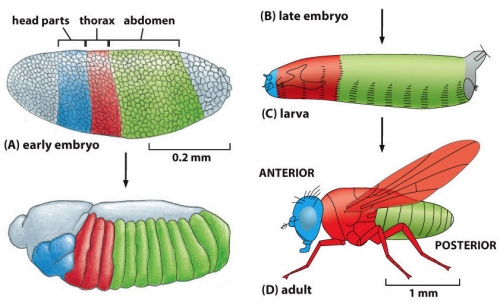


Figure 1-7 Life-cycle of *Drosophila melanogaster*. An adult female lays around 400 fertilized eggs (embryos). In the first few hours, the early embryo undergoes multiple rapid nuclear divisions until some 5000 nuclei accumulate in the unseparated cytoplasm (syncytium). The eggs hatch after 12-15 hours and the resulting larvae grow for about four days while molting twice. Then the larvae encapsulate in the puparium and undergo a four-day-long metamorphosis, after which the adults emerge. (adapted from StudyBlue)

Die Larve zeigt bereits ein Segmentierungsmuster und ist in Kopf, Thorax und Abdomen aufgeteilt. Die embryonalen Segmente entsprechen dabei den späteren Segmenten der Larve. Dadurch können Segmentierungs-Mutationen einfach beobachtet werden.



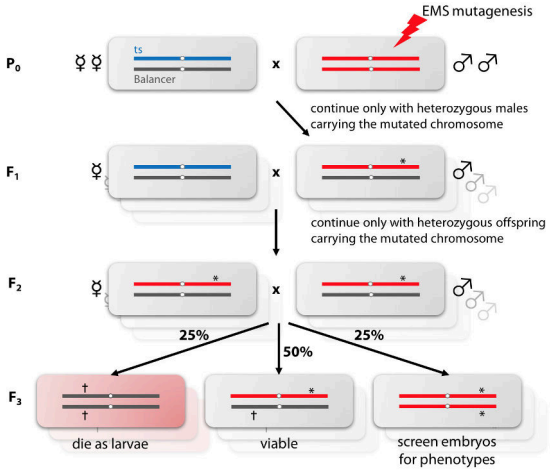
Ein wichtiger Faktor in der Entwicklung von *Drosophila* ist der Maternaleffekt. Die Mutter belädt das Ei mit Genprodukten (z.B. Bicoid-mRNA), welche die ersten Schritte der Embryonalentwicklung massgeblich beeinflussen. Entsprechend sind nur sehr wenige zygotische Gene für die ersten Schritte nötig und die meisten Mutationen beeinflussen die frühe Embryogenese nicht.
Weiterhin lassen sich viele Phänotypen wie Augenfarbe, Flügelform oder Borstenzahl einfach unter dem Binokular beobachten. Es gibt Sammlungen verschiedener Fliegenmutanten (*mutant libraries*) oder genetisch modifizierter Fliegen, die Forschern frei zugänglich sind.
Ein grosser Nachteil von *Drosophila* als Modellsystem ist, dass Eier oder Larven nicht eingefrohren und wiederbelebt werden können.

5.2 Balancer-Chromosomen

Balancer-Chromosomen tragen einen dominanten phänotypischen Marker. Gebogene Flügel können z.B. ein solcher Marker sein. Somit ist nach Kreuzugen ersichtlich, welche Fliegen dieses Chromosom tragen. Um die Möglichkeit eines *crossing over* auszuschliessen, enthalten Balancer-Chromosomen mindestens einen umgekehrten DNA-Abschnitt. Der Verlust der Homologie führt dazu, dass Zellen mit Rekombination sterben, weil zu viele Deletionen oder Duplikationen vorkommen. Viele Balancer-Chromosomen tragen ausserdem eine rezessive, letale Mutation. Auf diese Weise sterben auch Zellen mit zwei Balancer-Chromosomen. Durch all diese Tricks ist die Population der für das Balancer-Chromosom und die Mutation heterozygoten Fliegen stabil. Es gibt für jedes *Drosophila*-Chromosom ein entsprechendes Balancer-Chromosom.

5.3 Heidelberg-Screen zur Identifizierung Embryonalentwicklungsgenen

Chrisitane Nüsslein-Vollhard und Eric Wieschaus identifizierten die 15 grundlegenden Gene im sogenannten Heidelberg-Screen. Dabei mutagenisierten sie Fliegen der Parentalgeneration P_0 und kreuzten deren Nachkommen der Filialgeneration F_1. Die zweite Filialgeneration F_2 wurde wiederum gekreuzt. 25% der Fliegen der F_3-Generation waren homozygot für das mutierte Chromosom. Wenn also Fliegen ohne Balancer überlebten, war die Mutation nicht letal. Sonst war die Mutation embryonal oder larval letal. In der frühen Larve ist das Segmentierungsmuster bereits in der äusseren Schicht, der Cuticula, sichtbar. Die Gene sind nach ihrem Phänotyp in folgende Klassen aufgeteilt: Lücken-, Paarregel- und Segmentpolaritätsgene. War die Mutations eines Gens früh in der Entwicklung letal, konnte seine Funktion in späteren Stadien allerdings nicht untersucht werden.



5.4 Klonale Screens

Um das Problem der Letalität von Mutationen zu umgehen, kann sie nur in bestimmten Zellen des Organismus homozygot hervorgerufen werden. Dies wird über Rekombination in mitotischen (somatischen) Zellen erreicht. Man verwendet dazu das Hefe-Enzym Flp-Rekombinase, welches ortsspezifische Rekombination zwischen sogenannten FRT-Stellen verursacht. Diese Stellen werden künstlich in die Fliege eingeführt. Inzwischen existieren Fliegenlinien mit FRT-Stellen auf jedem Arm jedes der Hauptchromosomen (X, 2, 3). Auch die Flp-Rekombinase ist ein Transgen. Je nach dem wie viele solche Ereignisse vorkomen und wo sie stattfinden, kann Gewebe von homozygoten Chromosomenarmen bis zu fast komplett homozygot reichen. Um die Rekombination beobachten zu können, sind wieder sichtbare Marker auf dem selben Arm wie die Mutation im Spiel.
Vorteile von klonalen Screens sind, dass F_1-Screen für rezessive Phänotypen durchgeführt werden können und dass mithilfe der Flp-Rekombinase die Mutation gewebespezifisch aktiviert werden kann (z.B. nur in Augenzellen).

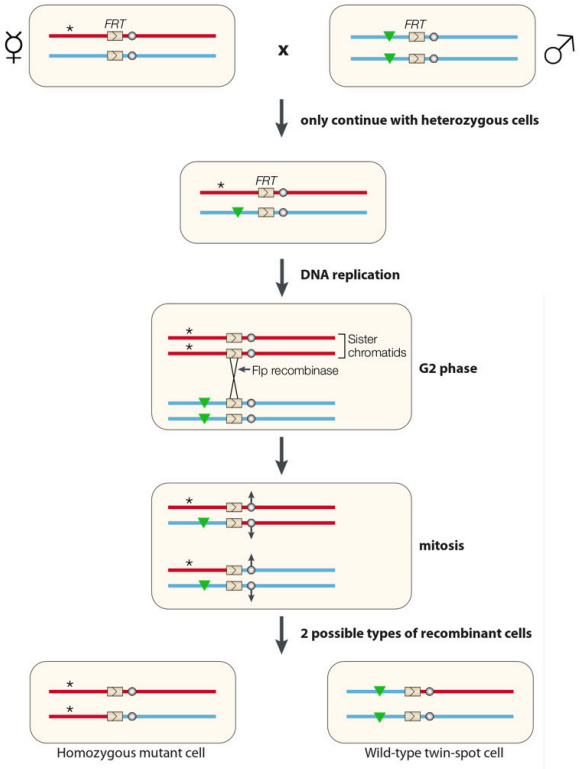
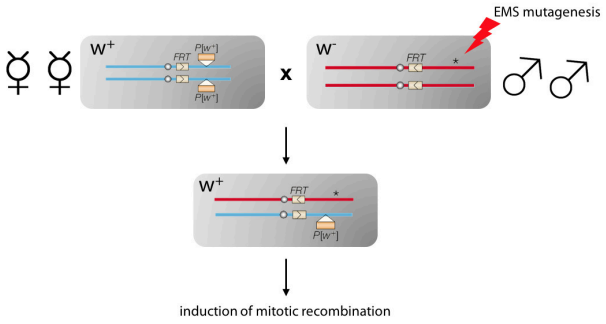


Figure 2-1 Crossing scheme for generating homozygous mutant clones using Flp/FRT-induced mitotic recombination. The green triangle indicates position of a GFP marker. (adapted from D. St Johnston, *Nat. Rev. Genet.*)

5.4.1 Gewebespezifische klonale Screens

Das Flp/FRT-System kann mithilfe eines gewebespezifischen Promotors erweitert werden. Ein Beispiel dafür ist das ey-Flp-System, in dem die Flp-Rekombinase mit dem Regulator des *eyeless*-Gens fusioniert ist und somit nur in der sich entwickelnden Augenscheibe exprimiert wird.
Das gesamte Genom kann mit vier Screens (einer pro Hauptchromosomenarm) untersucht werden. Da nur Weibchen zwei X-Chromosomen haben, wendet man dafür andere Tricks an.



5.5 Herstellung transgener Fliegen

Mit einer feinen Kanüle wird das Transgen in das posteriore Ende des Syncytium (mehrkernige Eizelle) injiziert, da sich dort Vorläufer der Keimzellen befinden. Dort muss es mit dem Fliegen-Genom rekombinieren, um stabil erhalten zu bleiben. Die transgene DNA enthält wieder ein Markergen. Oft wird das rote Augen verleihende *white*$^+$ verwendet. Gene werden in Fliegen nach ihrem mutanten Phänotyp benannt. *white*$^+$ ist dabei der Wildtyp (WT) und *white*$^-$ die Mutante.

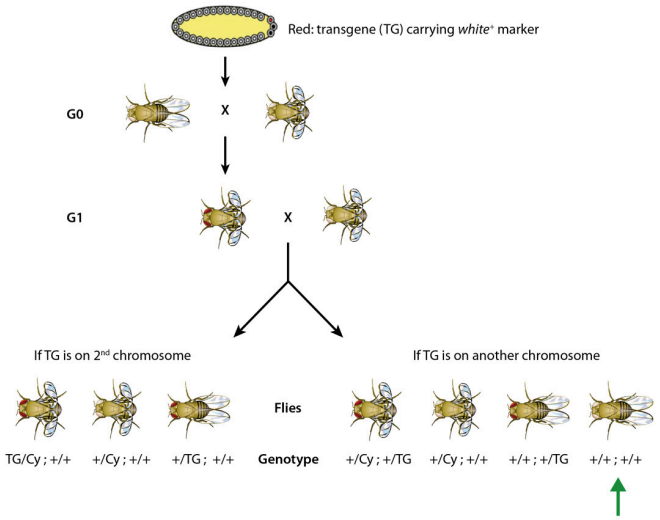


Figure 2-10 Crossing scheme to determine the insertion site of transgenes in *Drosophila*. After injecting the transgene into embryos, these embryos develop into adult flies (G0). G0 flies are crossed with flies containing a balancer chromosome, here, one for the 2nd chromosome, carrying the dominant marker *Cy* that confers bend up wings. In the G1 generation, flies with red eyes will appear, indicative of transgenic animals. To determine whether the transgene is on the 2nd chromosome, G1 flies are crossed again with flies containing the balancer. There are two different possibilities for the phenotypes of flies in the G2 generation (left or right), with only one phenotype that is different between both cases (green arrow). This is the one determining the location of the transgene. In both cases, both chromosomes (second and third) are indicated.

5.6 Herstellung transgener Fliegen mithilfe von Transposons

In *Drosophila* wird vorallem die Transposon-vermittelte Integration von Fremd-DNA mithilfe des P-Elements angewendet. Das P-Element enthält zwei für die Mobilisierung und Transposition wichtige terminale Wiederholungen. Es kodiert für das Enzym P-Transposase. Damit das Transposon nicht von alleine Springt, wird das P-Transposase-Gen zerstückelt und somit ein nicht-autonomes P-Element hergestellt. Damit das P-Element aber in das Genom intergriert werden kann, wird ein weiteres Plasmid mit der P-Transposase injiziert. Die Insertionsstelle ist im Allgemeinen zufällig, es gibt aber sogenannte *hot spots*, an denen P-Elemente besonders häufig integriert werden.

6 Entwurf einer genetischen Studie

6.1 Reduktionistische vs. genetische Forschungsansätze

In mechanistischen / reduktionistischen Ansätzen wird ein biologischer Prozess in Teilprozesse und physische Komponenten zerlegt und diese Einzelteile dann beobachtet. Genetische Studien untersuchen die Folgen eines nicht mehr funktionierenden Prozesses, um ihn zu verstehen.

6.2 Vorgehen bei einer genetischen Studie

1) Auswahl eines Modellorganismus zu dem darin ablaufenden Prozess. Dabei eignen sich bestimmte Organismen aufgrund ihrer Eigenschaften und der zu verfügung stehenden experimentellen Werkzeuge.

2) Auswahl eines Phänotyps, welcher Veränderungen des Prozesses veranschaulicht. Ein guter Phänotyp hat dabei eine hohe Spezifität. Er lässt sich also nicht oder kaum von anderen Prozessen verursachen. Ebenfalls ist wichtig zu beachten, wie aufwändig die Phänotypisierung bei der ausgewählten Kombination aus Organismus und Phänotyp ist. Man möchte einen möglichst effizienten Phänotyp. Ausserdem muss der Phänotyp auch plausibel sein, sprich genügend oft auftreten.

3) Erzeugung oder Identifizierung einer genetisch diversen Population. Da warten auf natürliche Mutationen zu lange dauert, kann man seine Population künstlicher Mutagenese unterwerfen und so übers ganze Genom verteilte Mutationen häufiger machen. EMS, UV und Transposons sind beliebte Mutagenesemethoden (siehe Kapitel 4.5). Man muss sich dabei bewusst sein, dass *gain-of-function*- und *loss-of-function*-Mutationen viel seltener sind als stille Mutationen.

4) Systematisches Untersuchen aller Individuen auf den ausgewählten Phänotyp.

5) Identifizierung der verantwortlichen genotypischen Veränderungen. Wurden Transposons verwendet, um die genetische Diversität zu erzeugen, kann man nach deren Sequenz im Genom suchen. Bei anderen Mutagenesemethoden muss aber nicht direkt das ganze Genom sequenziert werden. Zum Beispiel grosse strukturelle Veränderungen auf chromosomaler Ebene können durch Methoden wie BioNano Mapping identifiziert werden. Punkmutationen und Indels dagegen sind einfacher mithilfe von Komplementations-basierten Methoden zu finden.

7 RNAi & CRISPR/Cas

7.1 RNA-Interferenz

RNAi ist ein zellulärer Prozess, bei dem kleine regulatorische RNA-Moleküle die Translation von mRNA verhindern und so die Gen-Expression unterdrücken. Heutzutage weiss man, dass es drei unterschiedliche Pathways gibt: miRNA (*micro*), siRNA (*small inhibitory*) und piRNA (*piwi-interacting*). piRNA ist dabei momentan am wenigsten verstanden und wird darum in dieser Vorlesung nicht weiter behandelt. Die Hauptunterschiede zwischen miRNA und siRNA sind:

	miRNA	siRNA
Ursprung	endogene, nicht protein-kodierende Sequenzen	exogene Sequenzen, z.B. viralen Ursprungs
Komplementarität zu target RNA	begrenzt	perfekt
zelluläre Funktion	modulieren oft grosse Anzahl Gene	Verteidigung gegen fremde RNA; deswegen spezifisch

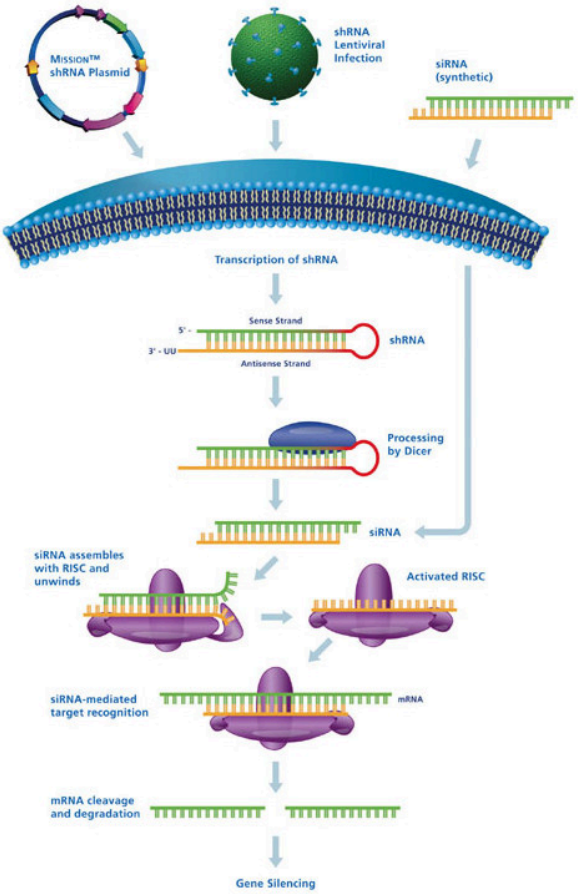
7.1.1 miRNA

miRNAs sind kurze, einsträngige und nicht-kodierende RNAs die auf post-transkriptioneller Ebene eine wichtige Rolle in der eukaryotischen Genregulation spielen. Sie inhibieren mRNA von Zielgenen durch Transkript-Destabilisierung und/oder translationale Hemmung. Dabei können sie über 100 verschiedene mRNAs inhibieren. Vermutlich sind miRNAs für für die Steuerung von über 60% aller menschlicher Gene verantwortlich. Weiterhin nehmen sie Einfluss auf Zellwachstum, Proliferation, Entwicklung, Differenzierung, Metabolismus und Apoptose.

miRNA-Gene befinden sich oft in Intronsequenzen neben den Genen, die sie regulieren. Als erstes werden sie in ein grosses Primärtranskript, die Pri-miRNA, transkribiert. Anschliessend wird eine Polyadenylkette und das 5'-Cap angehängt als auch alle Introns entfernt. miRNAs enthalten *inverted repeats*, welche das bilden von Haarnadelstrukturen erlauben. Diese werden vom Proteinkomplex Microprocessor erkannt. Er besteht aus den Proteinen Drosha und DGCR8. Ersteres schneidet den miRNA-Ste-Loop raus und isoliert somit die 60-80 Nukleotide lange pre-miRNA-Sequenz. Exportin 5 schleust die pre-miRNA danacht ins Zytoplasma. Das Protein Dicer bindet und schneidet so, dass eine doppelsträngige miRNA entsteht. Vom miRNA-Duplex wird zuletzt die nicht benötigte Sequenz entfernt, so dass die reife miRNA entsteht und an den RISC-Komplex (*RNA-induced silencing complex*) bindet. Ist die miRNA perfekt komplementär zu mRNA, bindet RISC letztere und die miRNA zerschneidet sie. Bei nicht perfekt komplementärer miRNA wird die mRNA lediglich gebunden (*translational block*). Die zweite Art der Regulation ist in Säugern weitaus häufiger.

7.1.2 siRNA

siRNAs sind ein wichtiger Verteidigungsmechanismus von Pflanzen, sind aber auch in anderen Eukaryoten zu finden. Gen-Knockdown-Experimente benutzen diese um spezifische mRNAs zu inaktivieren. Die doppelsträngige siRNA wird dabei entweder direkt in die Zelle eingeschleust, oder die dafür kodierende DNA wird über einen biologischen Vektor eingebracht. Die DNA-Sequenzen enthalten durch kurze Linker verbundene *inverted repeats*, welche das Bilden einer shRNA (*short hairpin*) erlauben. shRNA wird dann durch Dicer in siRNA umgewandelt. Die Vektor-Methode hat den Vorteil, dass das anvisierte Gen permanent ausgeschaltet wird, da kontinuierlich shRNA produziert wird.



Beim Design einer siRNA sind folgende Faustregeln zu beachten. Die 21bp lange Sequenz sollte in den ersten 50 - 100 Basenpaaren der mRNA liegen, zwei A am Anfang haben, mindestens 50% aus G-C bestehen und keine Möglichkeit für internte Hybridisierung bieten. Für wichtige Organismen sind inzwischen ausführliche siRNA-Bibliotheken verfügbar. Sonst gibt es Bioinformatik-Tools zur automatisierten Erstellung der Sequenzen.

7.1.3 Off-target Effekte

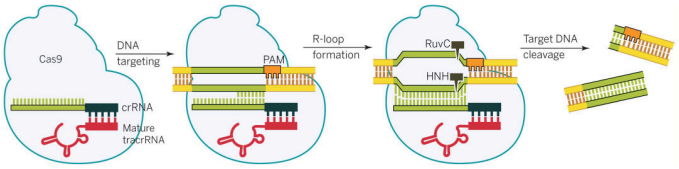
Selbst bei Verwendung von siRNAs besteht ein gewisses Risiko, dass die beobachteten Effekte durch unspezifische Inhibierung hervorgerufen werden. Dies kann vermieden werden durch besseres Design der siRNA oder durch Verwendung von mindestens drei siRNAs gleichzeitig. Wenn nun alle drei Experimente den selben Effekt zeigen, können *off-target*-Effekte höchstwahrscheinlich ausgeschlossen werden. Weiterhin ist ein Experiment mit irgend einer anderen siRNA zu empfehlen, damit für generelle Effekte durch Injektion kontrolliert werden kann. Ausserdem ist es üblich, die Konzentration der mRNA als weitere Kontrolle zu messen.

7.2 Das CRISPR/Cas9-System

Bakterielle Zellen können fremde DNA als Plasmide aufnehmen und weiterverwenden. Plasmide können aber auch Gefahren, wie aggressive Transposons, mit sich bringen. Ausserdem injizieren Phagen ihre DNA in Bakterien, wo Kopien des Phagen bis zum Platzen der Zelle produziert werden. Das CRISPR/Cas-System bietet die Möglichkeit, sich gegen solche Gefahren zu verteidigen.

7.2.1 Molekularer Mechanismus des CRISPR/Cas-Systems

CRISPR steht für *clustered regularly interspaced short palindromic repeats* und bezeichnet Abschnitte im Genom mit regelmässigen, oft palindromischen Repeat-Sequenzen. Zwischen den Repeat-Sequenzen sind von Phagen oder Plasmiden abstammende Spacer-Sequenzen angesiedelt. Sie bilden die Vorlage für das Verteidigungssystem. Spacer-Sequenzen werden bei Infektion aus der invasiven DNA rausgeschnitten und in das eigene Genom integriert. Es sind drei verschiedene Klassen von CRISPR-Systemen bekannt. Hier wird das viel genutzte CRISPR/Cas-System von *Streptococcus pyogenes* beschrieben.
Der gesamte CRISPR-DNA-Bereicht wird als einzige pre-crRNA transkribiert. Diese wird später in ihre Einzelteile zerlegt, welche wiederum mit tracrRNA (trans crRNA) hybridisieren. Dieses RNA-Dimer wird vom Cas9-Protein (*CRISPR associated*) gebunden. Es enthält ausserdem zwei Nukleasedomänen, welche die erkannte DNA durchtrennen. Die Erkennung basiert aus der Auftrennung des Ziel-DNA-Doppelstrangs und Bindung an die Spacer-DNA. Die PAM-Sequenz (5'-NGG-3' bei *S. pyogenes*) zeigt den Nukleasedomänen den Durchtrennungsort an.



Das bereits vergleichsweise einfache CRISPR/Cas-System von *Streptococcus pyogenes* kann weiter vereinfacht werden, indem crRNA und tracrRNA zu einem Molekül, der sgRNA (*singel guide*), vereint werden. Durch fortschreitende Optimierung der Cas9- und sgRNA-Sequenzen konnte die Effizienz auf über 80% gebracht werden.

7.2.2 Anwendung des CRISPR/Cas9-Systems in einem Geninaktivierungsexperiment

Das CRISPR/Cas9-System kann auf eine bestimmte Stelle des Genoms programmiert werden. Dabei sucht man im anvisierten Gen nach einer PAM-Sequenz. Die 20 5'-gelegenen Nukleotide werden dann in das sgRNA-Gen eingebaut. Vektoren für die sgRNA und Cas9 werden anschliessend in die Zielzelle eingeschleust. Meistens entstehen direkt homozygote Knockouts. Durch Erfahrung wurde festgestellt, dass es Unterschiede bei der Selektivität und Effizienz von unterschiedlichen Spacer-Sequenzen gibt und dass bestimmte Sequenzen die Effizienz erhöhen, wenn sie die PAM-Sequenz flankieren.

8 Zellliniengenetik

Forschung an Säugetieren und insbesondere Menschen bringt einige praktische und ethische Probleme mit sich. Deswegen werden häufig Zellkulturen verwendet.

8.1 Primäre vs Stabile Zellkulturen

Primäre Zellkulturen basieren auf durch Biopsie aus einem Säugetier gewonnenen Zellen. Sie überleben meist nur wenn ihr natürliches Umfeld mit Gewebe, Nährstoffen und Wachstumsfaktoren herrscht. Ausserdem ist die Anzahl Zellteilungen bei voll differenzierten Säugetierzellen limitiert (Seneszenz).

Stabile Zellkulturen basieren auf Zellen mit ausser Kraft gesetzter Zellzykluskontrolle und Abhängigkeit von der Umgebung und sind somit unsterblich (*immortalization*). Viele weitverbreitete Zelllinien, wie z.B. HeLa, wurden aus Tumoren gewonnen. Unsterblichkeit kann aber auch künstlich hergestellt werden. Entweder werden Seneszenzmechanismen ausgeschaltet oder die Zelle mit einer Krebszelle fusioniert (Hybridomazellen). Die genauen Mechanismen der Transformation sind noch nicht genau bekannt. Somit ist das Gelingen Glückssache.

8.2 Haltung von Säugetierzelllinien

Im Gegensatz zu typischen Labormikroorganismen benötigen Säugerzellen ein genau definiertes Wachstumsmedium mit Nähr- und Botenstoffen. Ausserdem sind Ausscheidungsprodukte der Zellen schnell toxisch für sie. Weiterhin wachsen viele der Zellen nur angeheftet auf mit bestimmten Proteinen und Polymeren beschichteten Oberflächen, welche das umgebende Gewebe simulieren sollen. Zellteilung geschieht nur ca. alle 24 Stunden. Zellkulturen sind bei Kontamination schnell überwachsen.

8.3 Genetik und Genomik bei Säugetierzelllinien

Säugetierzellen sind diploid. Somit werden rezessive Phänotypen nur bei Homozygotie sichtbar. Weil sie sich asexuell durch Zellteilung vermehren, ist es unmöglich durch Rückkreuzung homozygote Mutationen zu erzeugen. *Crossing over* findet bei mitotischen Zellen auch nicht statt.
Genome von Säugetieren sind weitaus grösser und komplexer als diejienigen von Mikroorganismen. Sie enthalten nicht proteinkodierende Regionen aus repetitiven Sequenzen und Gene sind durch Introns unterbrochen. Zwei Duplikationsereignisse haben ausserdem zu einer hohen Redundanz geführt.

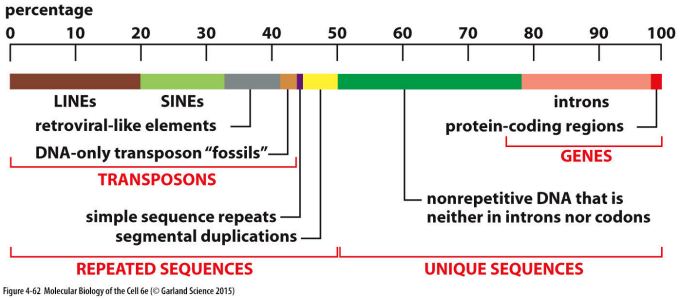


Figure 4-62 Molecular Biology of the Cell 6e (© Garland Science 2015)

Die in Kapitel 6.2 besprochene Art von genetischer Studie wird als *forward genetics* bezeichnet. Aufgrund der Eigenschaften von Säugetierzellen ist diese Art der Studie nicht praktisch. Treten *loss-of-function*-Mutationen auf, sind sie oft aufgrund der Redundanz nicht sichtbar. Selbst wenn eine Mutation einen Phänotyp generiert, ist die klassische Kartierung unmöglich. Stattdessen wird der sogenannte *reverse genetics*-Ansatz gewählt. Dabei inaktiviert man zuerst ein Gen und versucht dann die phänotypische Veränderung zu beobachten.

8.3.1 Transiente und stabile Überexpression von Genen

Das Einführen und dann Exprimieren von Genen in eine Zelle ist eine der klassischen Methoden der reversen Genetik. Dies kann z.B. über Plasmide geschehen. Eine bestimmte Methode verwendet dabei Lipofectamin, mit dem das Plasmid umhüllt wird und das Liposom dann zur Nährlösung gegeben wird. Dort fusioniert es mit der Plasmamembran. Das Plasmid wandert in den Zellkern und wird dort transient exprimiert. Die bedeutet, dass es nicht in das Genom integriert wird, sondern direkt als Plasmid transkribiert wird.
Will man Gene permanent in eine Zelle einführen, werden retrovirale Transduktionssysteme verwendet. Ein Retrovirus intergriert das Gen zusammen mit einem Marker (oft GFP oder Resistenzgene) und seinem eigenen Genom in das Genom der Zelle.

8.3.2 RNAi und CRISPR/Cas9

RNAi ist die wohl einfachste Methode, um mit reverser Genetik einen *loss-of-function*-Phänotyp zu beobachten. CRISPR/Cas9 ist allerdings noch effizienter und schaltet meist beide Kopien eines Gens aus.

8.4 Haploide Zelllinien

Die von einem Laukämiepatienten stammende HPA1 Zelllinie besitzt bis auf Chromosome 8 und 15 nur je eine Kopie. Sie ist also fast haploid und unter Laborbedingungen stabil. Um überleben zu können, mussten aber einige Anpassungen geschehen. Z.B. sind HPA1 Zellen wesentlich kleiner.
Eizellen können durch induzierte Zellteilung vor der Befruchtung auch in haploide Zelllinien verwandelt werden.

9 Genetische Vielfalt beim Menschen: SNPs

„*The capacity to blunder slightly is the real marvel of DNA. Without this special attribute, we would still be anaerobic bacteria and there would be no music.*"
— Lewis Thomas

Genetische Vielfalt entsteht bei verschiedenen Spezies auf unterschiedliche Art und Weise. Die hier besprochenen Mechanismen sind auf den Menschen bezogen. Je grösser die evolutionäre Entfernung, desto wenig sind sie auf andere Spezies anwendbar.

9.1 Grösse des Genoms

Das menschliche Genom ist rund 3 Milliarden Basenpaare gross. Obwohl zwei nicht miteinander verwandte Menschen immernoch etwa 99.9% des Erbguts teilen, sind im Schnitt immer noch ca. 6 Millionen Basenpaare verschieden.

9.2 Single Nucleotide Polymorphisms

Es gibt etwa 40 Millionen Stellen, an denen sich das menschliche Genom zwischen Individuen besonders oft unterscheided. SNPs sind die häufigeste Art von genetischer Variation. Dabei ist wichtig zu beachten, dass sie keine Beschädigung der DNA im Sinne von *basepair mismatch* sind. Ist das eine Nukleotid anders, ist auf dem anderen Strang das komplimentäre dazu zu finden. Eine Einzelbasensubsitution wird nur als SNP bezeichnet, wenn sie in mindestens 1% der Bevölkerung auftritt. Sonst heisst sie private Punktvariante. SNPs sind sind in der Regel über 100'000 Jahre alt und wurden im Laufe der Zeit mehrmals durch *crossing over events* zwischen homologen Chromosomen ausgetauscht. Somit folgt die Verteilung bestimmten statistischen Mustern.

SNPs starten als private Punktmutationen in einer Keimbahnzelle. Pro Generation treten rund 30 neue davon auf und werden vielleicht an die nächste Generation weitervererbt. Bei einem Elternpaar mit zwei Kindern besteht eine Wahrscheinlichkeit von 25%, dass die Mutation nicht in den Kindern vorkommt. Dies gilt für jede weitere Generation ebenfalls. Wenn aber durch Zufall oder Selektion eine Punktvariante sich in der Population ausbreitet, gelangt sie nach einigen hundert oder tausend Generationen an einen Punkt, an dem die Wahrscheinlichkeit, dass sie verloren geht, stark abfällt. Dabei wird sie durch *crossing over events* zusammen mit naheliegenden SNPs vererbt und es entsteht das Co-Vererbungsmuster. Durch Drift kann die Variante entweder wieder verschwinden oder irgendwann in 100% der Population vorhanden sein. Dann spricht man von einem fixiertem SNP. Man geht davon aus, dass von der Entstehung bis zur Fixierung eines SNPs ca. 250'000 Jahre vergehen. Schneller kann es gehen, wenn die neue Variante einen evolutionären Vorteil bietet. Das Lactase-Gen *LCT* hat eine Variante, die es Menschen nach dem Alter von 3-4 Jahren immer noch Lactose abzubauen erlaubt. Verantwortlich dafür sind zwei Punktmutationen, welche die Abschaltung des Gens verhindern. Vor einigen tausend Jahren bot diese Variante in Nordeuropa einen wesentlichen Vorteil, was es ihr erlaubte sich über fast die gesamte Population auszubreiten.

9.3 Haplotypen

Crossing over events sind ein gewollter Prozess, der immer wieder neue Kombinationen von vorhandenen Varianten generiert. Nahe gelegene SNPs werden besonders oft miteinander vererbt (*linkage disequilibrium / LD*). Orte, an denen besonders oft *crossing over events* stattfinden, werden *hot spots* genannt. Somit werden oft die selben SNPs zusammen vererbt. Diese Blocks nennt man LD-Blocks. Sie haben als Konsequenz, dass genetische Variationen als Haplotypen auftreten. Es reicht daher meist aus, einige wenige SNPs zu bestimmen, um den gesamten LD-Block zu kennen.

10 GWAS am Menschen

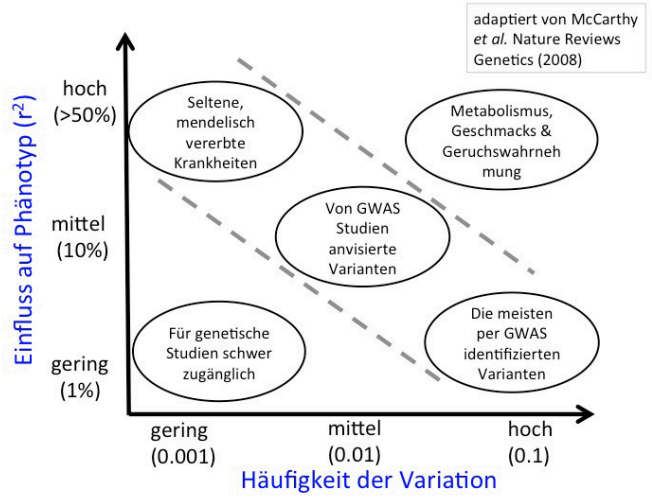
GWAS steht für genomweite Assoziationsstudie bzw. genome-wide association study und ist eine Methode zur Untersuchung von genetischen Faktoren und ihres Einflusses auf Phänotypen. Phänotypen, die von einem einzelnen Gen beeinflusst sind, sind schon länger genetisch erklärbar. Die viel häufigeren quantitativen Phänotypen hingegen sind viel schwieriger mit ihren mehreren Genen zu assoziieren.

10.1 Kandidaten/Gen-Studien

Nach der Einführung von automatisierten DNA-Sequenziertechniken wurde begonnen, sogenannte Kanditaten-Gen-Assosiationsstudien durchzuführen. Dabei wir durch Überlegungen ein Gen bestimmt und darauf getestet, ob es mit einem bestimmten Phänotyp assoziiert ist.

10.2 Seltene Krankheiten vs. common disease – common variant

Seltene Krankheiten treten bei weniger als 1 von 2000 Personen auf und haben ihre Ursache meistens in einer einzelnen Variation. Es können trotzdem zwischen verschiedenen Betroffenen der selben Krankheit verschiedene Mutationen auftreten.
Weitverbreitete Krankheiten hingegen sind laut der *common disease – common variant* Hypothese auf weitverbreitete Variationen zurückzuführen. GWAS wurden entwickelt, um diese Hypothese zu testen und solche Zusammenhänge zu finden.



10.3 GWAS

GWAS untersuchen statistische Zusammenhänge zwischen genotypischer und phänotypischer Variation. Dabei werden keine Annahmen über involvierte Gene getroffen. Man wählt lediglich einen zu untersuchenden Phänotyp. Die Genotypisierung der Studienteilnehmer wird mithilfe von Microarrays, die rund 1 Million SNPs messen und die LD-Blocks bestimmen, durchgeführt. Für jeden SNP wird anschliessend ein Assoziationstest durchgeführt. Für quantitative Phänotypen wird eine Regressionsgerade mit der Methode der kleinsten Quadrate berechnet. Mit der Steigung β, der Standardabweichung der Steigung σ_β und der Anzahl Teilnehmer n kann über die Student-t-Verteilung *tcdf* der p-Wert berechnet werden. Der p-Wert ist die Wahrscheinlichkeit, dass man per Zufall den beobachteten oder einen noch extremeren Wert der Steigung erhält.

$$p = 2 * tcdf\left(-\left|\frac{\beta}{\sigma_\beta}\right|; n\right)$$

Für qualitative Phänotypen verwendet man andere Tests, die auf logistischer Regression basieren. Das primäre Resultat einer GWAS ist ein p-Wert für jeden getesteten SNP, die dann zur Ansichtlichkeit geplottet werden können.

10.4 Herausforderungen und Annahmen in GWAS

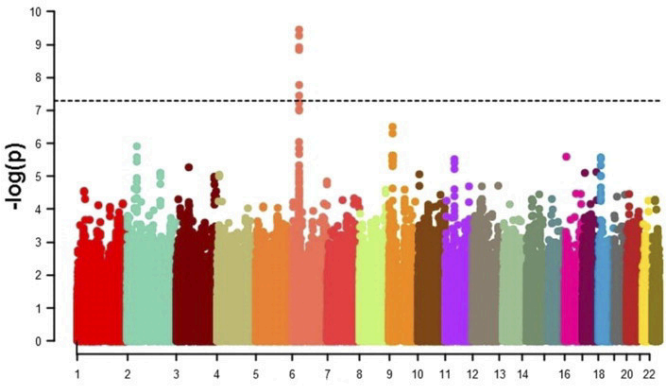
- Die erste Annahme ist die Normalverteilung der Phänotypen. Ist dies nicht der Fall, kann eine Transformation angewendet werden. Allerdings sollte man sich überlegen, ob es nicht eine passendere Messgröse für den Phänotyp gibt. Ausserdem sollte die Streuung (Varianz) der Phänotypen für die unterschiedlichen Genotypen gleich sein (Homoskedastizität).
- Bei ungenauen Messungen oder zeitlich schwankenden Phänotypen, muss man Massnahmen treffen um statistisches Rauschen zu eliminieren. Dies kann zum Beispiel über Mittelung mehrere Messungen geschehen.
- Für den Phänotyp beeinflussende Faktoren (Covariablen) sollte unbedingt korrigiert werden.
- Beobachtungen sollten unabhängig voneinander sein. Man sollte also nicht zwei Messungen beim selben Teilnehmer vornehmen und dann n erhöhen.
- Ein dritter Faktor könnte sowohl mit den Genotyp, als auch mit dem Phänotyp korreliert sein und so eine indirekte Assoziation verursachen. Man nennt diese Assoziationen *population stratification*. Dabei sind die Teilnehmer nicht aus einer genetisch homogenen Population, sondern einer distinkten Unterpopulation (z.B. definiert durch Ernährungsart, körperliche Aktivität, medizinische Versorgung, etc.). Da dies aber schwierig komplett zu verhindert ist, verwendet die gängige EigenStrat-Methode die sogenannte *principal component analysis* (PCA) um für solche Effekte als Covariablen zu korrigieren.

10.5 Bonferroni-Korrektur und genomweite Signifikanz

Da die Bonferroni-Korrektur eine Signifikanz von 10^{-8} $(= \frac{0.01}{1'000'000})$ verlangt, hat man sich auf einen *cutoff* von $10^{-7.5}$ geeinigt. Dies gilt auch, wenn wesentlich mehr als 1 Million SNP verwendet werden. Überschreitet ein SNP diesen Signifikanz-Wert, gilt der nächstgelegene SNP auch schon als signifikant bei $p = \frac{10^{-7.5}}{2}$ und so weiter.

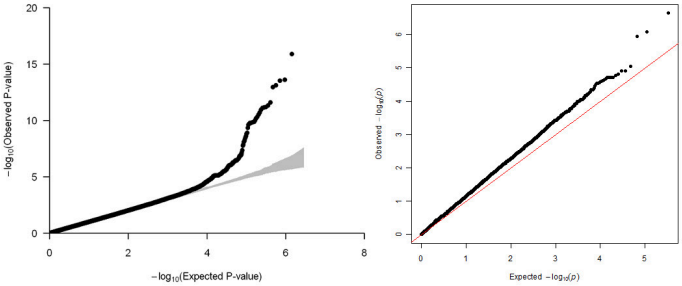
10.6 Manhattan-Plots

Die x-Achse gibt die Position im Genom in Form von Chromosomennummer und Basenpaarposition an, während die y-Achse den p-Wert der Assoziation mit dem Phänotyp angibt. Letztere ist dabei oft logarithmisch. Mit einer gestrichelten Linie kann der Signifikanz-Schwellenwert angezeigt werden.



10.7 QQ-Plots

Bis auf eine relativ kleine Anzahl SNPs, sollte die Assoziation dem Zufall folgen. Ist dies nicht der Fall, liegt wahrscheinlich ein Fehler in der Analyse vor. Um dies grafisch darzustellen, verwendet man QQ-Plots. Dafür sortiert man alle beobachteten und alle erwarteten p-Werte der Grösse nach und plottet die entstandenen Wertepaare in einem Scatter-Plot. Die Achsen sind auch hier logarithmisch. Die folgende Abbildung dient zur Veranschaulichung: Der linke Graph zeigt einen Abschnitt mit perfekter Korrelation, während einige Dutzende SNPs im oberen rechten Bereich einen klar niedrigeren p-Wert als erwartet haben. Der rechte Graph zeigt eine Abweichung von der erwarteten Geraden und weisst so auf systemische Probleme hin.



10.8 Genomic control

Da QQ-Plots selten so perfekt ausfallen wie im linken Beispiel weiter oben dargestellt, werden kleinere Abweichung von der Ideallinie mit *genomic control* korrigiert. Der Parameter λ gibt dabei die Qualität der GWAS an. Ein λ von 1 bedeutet, dass keine Korrektur nötig war. Im Allgemeinen sind λ's von 0.95-1.05 akzeptabel.

10.9 β und r^2 für quantitative Phänotypen

β gibt die Steigung der Regressionsgerade an. r^2 ist die *explained variance*, sprich der Anteil der phänotypischen Varianz, welche durch den Genotyp an diesem SNP erklärt wird. r^2 kann dabei von 0 (erklärt nichts) bis 1 (erklärt alles) reichen. Bei r^2 von über 0.5 spricht man meistens schon von Mendel'schen Assoziationen.

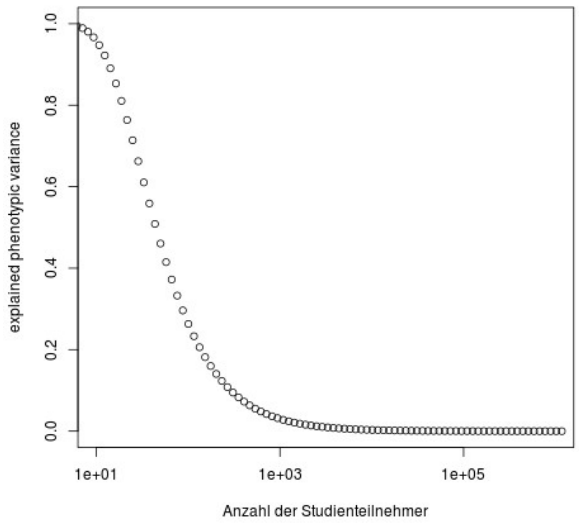
10.10 Odds ratio für qualitative Phänotypen

Das OR gibt an, wie stark der Genotyp das Verhältnis zwischen Teilnehmern mit und Teilnehmern ohne Phänotyp beeinflusst. Da bei qualitative Phänotypen meist logistische Regression genutzt wird, erhält man den Regressionskoeffizienten β_L. Dabei gilt dann $\text{OR} = e^{\beta_L}$.

10.11 Teilnehmerzahl, erklärte Varianz, statistische Signifikanz, phänotypisches Rauschen

Teilnehmerzahl n, erklärte Varianz r^2, statistischer Signifikanz p. r und n kompensieren sich gegenseitig.

$$p = 2 * tcdf\left(-\sqrt{n\frac{r^2}{1-r^2}}, n\right)$$



Weiterhin ist die phänotypische Varianz nicht nur von biologischen Faktoren abhängig, sondern auch von Faktoren wie Umwelt, statistisches Rauschen, etc. Da diese Varianzen alle additiv sind, sollte man versuchen diejenigen von nicht biologischen Faktoren zu minimieren, um den relativen Einfluss der SNPs zu erhöhen.

10.12 Ablauf von GWAS

1) Auswahl eines Phänotyps, der einfach und präzise zu messen ist und den biologischen Prozess gut beschreibt.
2) Abschätzung der Erblichkeit und der benötigten Anzahl Teilnehmer
3) Teilnehmer-Rekrutierung und Messung von Phänotyp und Covariablen
4) Genotypisierung
5) Qualitätskontrolle der SNPs
6) Imputation (Ersetzen fehlender Datenpunkte) von zusätzlichen SNPs (optional)
7) Bevölkerungsstruktur bestimmten mit PCA
8) Bestimmung und Korrektur von Covariablen
9) GWAS-Analyse
10) Korrektur von p-Werten durch *genomic control*
11) QQ- und Manhattan-Plots

10.13 Replikation

Da GWAS derart komplex und die phänotypischen Effekte oft so gross sind, wird bei grossen Effekten ($r^2 > 30\%$) eine zweite, unabhängige Studie verlangt. Idealerweise ist dabei die Genotypisierungsplattform, die Bevölkerungsgruppe der Teilnehmer und die Messart des Phänotyps unterschiedlich zur ersten Studie. Da nun ein Kandidatengen untersucht wird, ist die Bonferroni-Korrektur nicht mehr nötig und man kann deutlich weniger Teilnehmer verkraften.

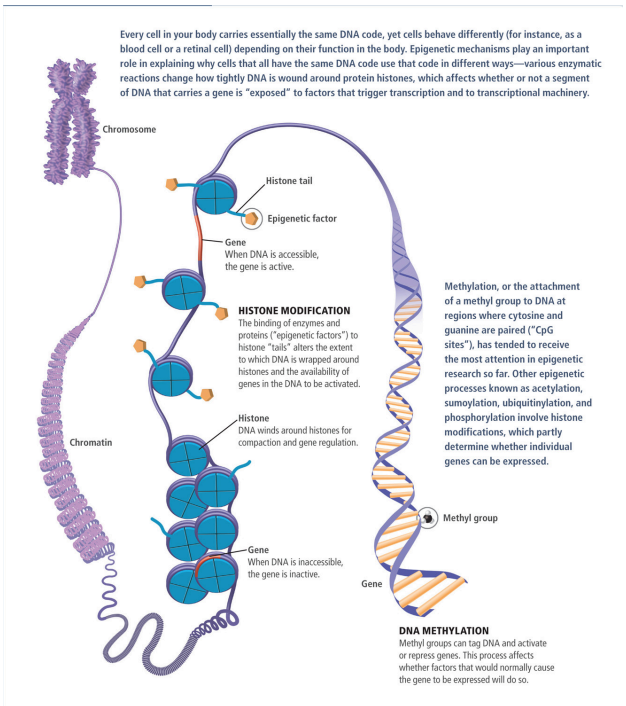
10.14 Kausal- vs. Proxy-SNP

Kausale SNPs verändern den Phänotyp, wenn sie eine andere Variante haben. Proxy-SNPs sind dagegen nur starkem LD unterworfen, werden also zusammen mit einem Kausal-SNP vererbt. GWAS können zwischen diesen zwei Arten von SNP nicht unterscheiden.

11 Epigenetik

Zwei Zellen eines Individuums haben die gleiche Genomsequenz. Die Unterschiede sind durch unterschiedliche Expressionsmuster zu erklären. Die Epigenetik (Epi- = hinzu / darüber hinaus) befasst sich mit Mechanismen der Vererbung, die nicht in der DNA sondern in Expressionmustern von Genen kodiert sind. Die Expression wir dabei durch reversible chemische Modifikationen der DNA oder der Histone verursacht. Diese Modifikationen werden dann an die Tochterzelle weitergegeben.
Teilweise werden auch generell langfristig stabile Veränderungen der Genregulation als epigenetisch bezeichnet. Gemeint sind z.B. Histonmodifikations-bedingte, lokale Veränderung der Chromatin-Packungsdichte oder RNA-abhängige Effekte, von denen nicht unbedingt gezeigt wurde, dass sie von Mutter- zu Tochterzelle vererbt werden. Solche Mechanismen werden teilweise auch als nicht-genetische Vererbung bezeichnet. Transgenerationelle epigenetische Vererbung könnte ein evolutionärer Vorteil sein, da sie zu schnellerer Anpassung führen könnte. Z.B. beeinflussen die Ernährungsgewohnheiten der Grosseltern die Lebenserwartung deren Enkelkinder (Överkalix, Schweden) oder die Unterernährung der Grossmutter während der Schwangerschaft führt zu einem reduzierten Geburtsgewicht des Kindes und zu Übergewicht bei neugeborenen Enkelkindern (Hongerwinter 1944/45).
Epigenetik ist kein, wie oft fälschlicherweise dargestellt, Beispiel für Lamarck'sche Evolution. Sie bietet lediglich einen Mechanismus, der parallel zur klassischen Darwinistischen Evolution die reversible Vererbung ermöglicht.

11.1 Molekulare Mechanismen



11.1.1 Chromatinstruktur

DNA in eukaryortischen Zellen ist eng auf Histonproteine aufgewickelt. Diese Kombination wird auch als Nucleosom bezeichnet. Generell gilt, dass je enger die DNA aufgewickelt ist, desto unwahrscheinlicher ist die Transkription von sich dort befindenden Genen. Dicht gepackte Regionen werden als Heterochromatin bezeichnet, während weniger dichte Euchromatin heissen. Veränderung dieser Packungsdichte wird als *chromatin remodelling* bezeichnet.

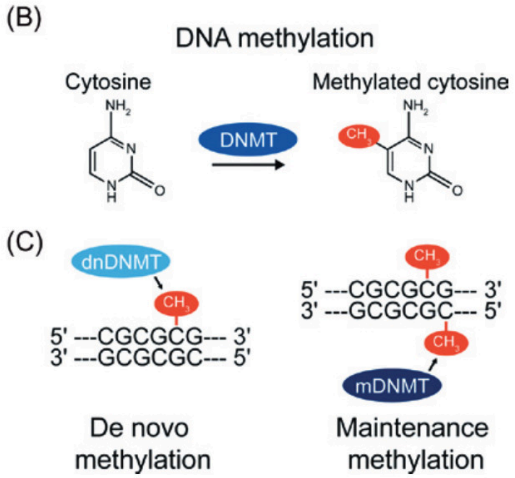
11.1.2 Histonmodifikation

Der Hockeypuck-ähnliche Kern des Nucleosoms besteht aus einem Octamer von Histonproteinen (je zwei Kopien von H2A, H2B, H3 und H4). Alle diese Proteine haben ausser dem zentralen auch einen flexiblen Teil, der *histone tail* genannt wird. Die Aminosäuren in diesen sind oft post-translational modifiziert. Modifikationen können mono-, di- und tri-Methylierung von Lys, Methylierung von Arg, Phosphorylierung von Ser, Thr und Tyr als auch Acetylierung von Lys sowie Ubiquitinierung und SUMOylierung sein. Es werden weiterhin immer weider neue Modifikationen entdeckt und jede von ihnen hat ein zugehöriges Protein, dass sie verursacht oder rückgängig macht.
Obwohl viele Aspekte der Histonmodifikation noch aktiv erforscht werden, sind folgende Tatsachen etabliert:

- Histonmodifikationen wirken lokal. Eine Ausnahme bietet die Inaktivierung eines gesammten X-Chromosoms in weiblichen Zellen.
- Histonmodifikation und deren Effekte sind reversibel. Dies geschieht mithilfe von Enzymen, sogenannten *readers* und *writers*.
- Acetylierte Lys-Seitenketten sind ungeladen und schwächen die Bindung zwischen DNA und Histonen, was eine lockerere Packungsdichte zur Folge hat.
- Die Effekte von Lys-Seitenketten-Methylierung hängen stark von deren Methylierungs-Grad und der betroffenen Aminosäure ab. Die Methylierungsmuster werden vermutlich von spezifischen Enzymen erkannt.

11.1.3 DNA-Methylierung

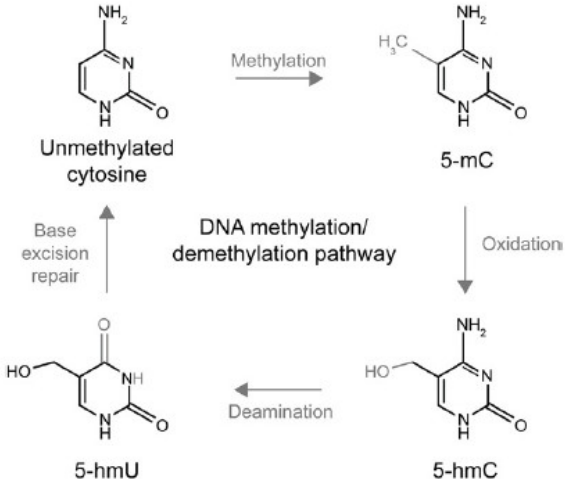
Epigenetische DNA-Methylierung beschreibt die spezifische Modifikation an der 5-Position des Cytosin-Basisrings. Vorallem solche vor einem Guanin werden methyliert. Dabei spricht man häufig von einem CpG-Dinukleotid. CpGs sind meist methyliert. Unmethyliert kommen sie vorallem in *CpG islands* vor. Diese Inseln sind besonders häufig in Promotoren (ca. in 60%) und in von Retro-Transposons abstammenden Abschnitten. Letztere sind von Zelltyp zu Zelltyp weitgehend gleich stark methyliert, während Promotoren starke Unterschiede aufweisen. Diese Sequenzen werden als *differentially methylated regions* (DRM) bezeichnet.
Die Methylgruppen zeigen in die *major groove* der DNA, wo sie die Bindung von Proteinen wie Transkriptionsfaktoren beeinflussen. Es gibt aber auch CpG-bindende Proteine, die selber das Binden von anderen Proteinen verhindern oder födern.



Die Methylierung wird von Methyltransferasen (DNMTs) ausgeführt, wobei der Cofaktor S-Adenosyl-Methionine (SAM) als Methylquelle dient. Dieser wiederum ist von Methyldonoren wie Folsäure abhängig. *De novo*-Methylierung wird von DNTM3A und DNTM3B ausgefürt. *Maintenance*-Methylierung geschieht auf dem Tochterstrang bei der DNA-Replikation und wird von DNMT1 ausgeführt.
Die Methylierung von CpGs in regulatorischen Sequenzen führt im Allgemeinen zu reduzierter Transkription (*transcriptional silencing*). Ein höherer Grad der Methylierung verursacht dabei höhere Transkriptionsreduktion.

11.1.4 DNA-Demethylierung

Methylierte Cytosine sind sehr stabil. Sogenannte *ten eleven translocation*-Enzyme (TETs) können aber offenbar DNA dynamisch methylieren. Dabei wird oxidiert von Methyl-Cytosin zu Hydroxy-Methyl-Cytosin und anschliessend deaminiert zu Hydroxy-Methyl-Uracil, welches durch DNA-Reparaturmechanismen als beschädigt erkannt wird und durch ein unmethyliertes Cytosin ersetzt wird. Eine weitere Art der Demethylierung ist die Verhinderung von Methylierung von Tochtersträngen nach der DNA-Replikation.



11.1.5 Nicht-kodierende RNAs

Ob ncRNAs wie siRNA, micro RNA oder small nuclear RNA an epigenetischer Mechanismen beteiligt sind, ist noch nicht genau klar. Zellen mir deaktivierter RNAi zeigen allerdings mit epigenetischer Regulation assozierte Merkmale.

11.2 Agouti-Maus

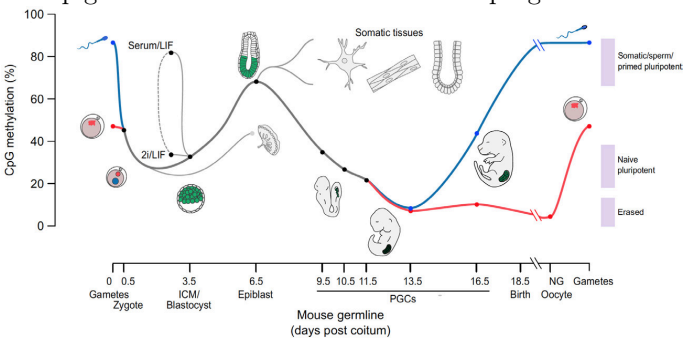
Wildtyp-Mäuse haben gelbliche Haare und an der Körperoberseite dunkelbraune bis schwarze Haarspitzen. Dieses Muster kommt von der Koordination von Haarwachstum mit dem Agouti-Gen. Es produziert ein parakrines Signal, welches Haarfolikel zum Wechsel von dunklem auf gelbes Haar veranlasst. Das WT-Allel wird als A und das mutante, nicht funktionstüchtige Allel als a bezeichnet. a/a-Mäuse haben dunkles bis schwarzes Fell.
Das A^{vy} Allel besitzt upstream des Agouti-Promotors eine als IAP (*intracisternal A particle*) bezeichnete Retrotransposonsequenz, welche einen kryptischen, sprich epigenetisch inaktivierten Promotor enthält. Ist er aber aktiv, wird das Agouti-Gen dauerhaft exprimiert. Das Fell bleibt komplett gelb und der Effekt in anderen Zellen führt zu Fettleibigkeit und anderen Veränderungen. Homozygot ist das Allel sogar lethal.
Die IAP Sequenz enthält CpGs, deren Methylierung den kryptischen Promotor deaktiviert und die Fellfarbe vom Wildtyp verursacht. Die Fellfarbe ist also ein Indikator für den Methylierungsgrad der IAP Sequenz. Interessant ist, dass Fleckenmuster über die Lebensdauer der Mäuse stabil sind. Die Methylierung wird also früh bestimmt und dann weitervererbt. Mütter mit fast ausschliesslich gelber Fellfarbe haben verstärkt Nachkommen mit nur gelbem Fell, während Wildtyp-ähnliche Mütter mehr komplett dunkle Nachkommen haben. Epigenetische Markierungen werden also während der Meiose nicht gelöscht. Ausserdem hängt der Methylierungsgrad der Nachkommen und die Vererbbarkeit von der Versorgung mit Methyl-liefernden Vitaminen während der Schwangerschaft ab.

11.3 Genomic Imprinting

In den meisten Fällen sind beide Kopien eines Gens aktiv, sprich je eine vom Vater und eine von der Mutter. Es gibt aber auch Gene, bei denen nur eine Kopie, z.B. die der Mutter, exprimiert wird. Man redet dabei von einem geprägten Gen (*imprinted gene*). Es sind vermutlich weniger als 1% aller Gene geprägt. Es scheint aber eine ganze Reihe Gene zu geben, bei denen zumindest geringe elternabhängige (*parent of origin*) Unterschiede auftreten. Der Grad der Prägung kann sich während der Entwicklung veränderun und sogar in die andere Richtung umschlagen. Die Mechanismen der Genprägung sind vorallem DNA-Methylierung. Es gibt aber auch Anzeichen, dass andere Mechanismen involviert sein könnten. Die Prägung kann unter anderem auch davon abhängen, ob sich das Gen in einer männlichen oder weiblichen Keimzelle befindet.
Die evolutionären Gründe für solch elterliche Einflüsse werden von der *parental conflict*-Hypothese erklärt. Die Mutter investiert gezwungenermassen mehr Energie als der Vater in den Nachwuchs. Für die Mutter ist es von Vorteil, wenn sie selber länger überlebt und mehr Nachwuchs mit verschiedenen Vätern zeugen kann. Im Interesse des Vaters ist es aber, dass sein Nachwuchs so schnell wie möglich, auch auf Kosten der Mutter, wächst. Tatsächlich stehen viele der geprägten Gene im Zusammenhang mit Wachstum und Metabolismus.

11.4 Epigenetisches Profil der Keimbahn

In Spermien werden 90-99% der Histone durch Protamine ersetzt. Gewisse HPTMs bleiben also erhalten. Es gibt auch einige Regionen im Genom, in denen Methylierung erhalten bleibt. Das Genom ist stark komprimiert, um die DNA in den kleinen Raum zu packen. Spermine können allerdings non-coding RNAs, wie miRNAs, tRFs, piRNAs, IncRNAs oder mRNAs, zur Eizelle transportieren und so Informationstransfer ermöglichen. In Oozyten sind die epigenetischen Prozesse ähnlich zu denen in somatischen Zellen. Die epigenetischen Muster aber werden reprogrammiert.



12 Krebsgenomik

Ein Tumor beschreibt die unkontrollierte Teilung von Zellen und das damit kommende Wachstum von Gewebe. Bösartige (maligne) Tumore werden dabei auch als Krebs bezeichnet. Gutartige (benigne) Tumore führen kein invasives Wachstum von Geweben aus. Der Entstehungsort eines Tumors wird als Primärtumor bezeichnet. Sekundäre Tumore (Metastasen) können in anderen Körperregionen gebildet werden, wenn Zellen eines Tumors die Basalmembran durchbrechen. Auch benigne Tumore können allerdings durch Druck auf benachbartes Gewebe Schaden anrichten. Maligne Tumore können wie folgt klassifiziert werden:

- Karzinome: Von Epithelzellen abstammend (80-90% aller menschlichen Tumore)
- Sarkome: Von Muskel-, Knochen- oder Bindegewebezellen abstammend
- Leukome: Von hämopoetischen (d.h. Blut-) Zellen abstammend
- neurologische Tumore

Alle Krebsarten basieren auf Mutationen. Diese gehen auf Replikationsfehler, chemische Karzinogene, ionisierende Strahlung, DNA-Viren, Retroviren oder Bakterien zurück. Dabei sind in den meisten Fällen mehrere Mutationen verantwortlich. Die betroffenen Gene können in Protoonkogene und Tumorsuppressorgene eingeteilt werden. Gesundes Zellwachstum kommt von einem Gleichgewicht zwischen Aktivierung und Inhibition. Des weiteren haben manche Menschen eine genetische Veranlagung für erhöhtes Krebsrisiko durch mehr Mutationen bei der Zellteilugng oder die Reaktion auf entstandene Mutationen.

12.1 Protoonkogene

In gesunden Zellen fördern Protoonkogene die Zellteilung und beugen den Zelltod vor. *Gain-of-function*-Mutationen können deren Aktivität steigern, wodurch sich die Zellen schneller teilen oder nicht mehr sterben. Durch diese Mutationen werden Protoonkogene zu Onkogenen. Wichtige Onkogene sind:

- c-Ha-Ras, c-Ki-Ras, c-N-Ras: GTPasen in Zellproliferations-Pathways
- c-Raf: Serin/Threoninkinase
- c-Jun, c-Fos, c-Myc: Transkriptionsfaktoren
- c-Src: cytoplasmische Tyrosinkinase
- c-Sis: Wachstumsfaktoren
- c-ErbB, HER2: Wachstumsfaktorrezeptoren
- BCI-2: Apoptose-Inhibitor

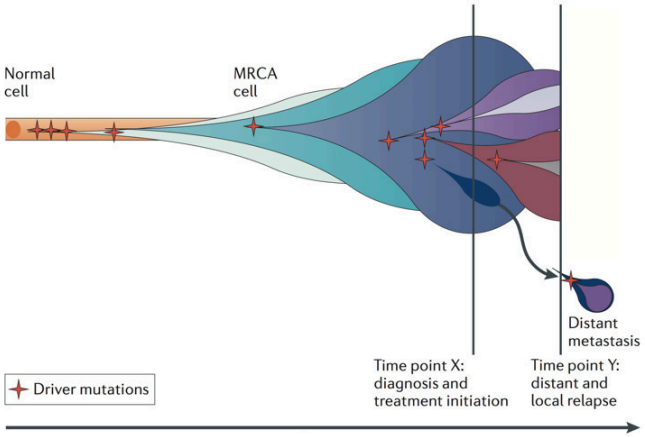
12.2 Tumorsuppressorgene

In gesunden Zellen kodieren Tumorsuppressorgene für Proteine, die Zellwachstum kontrollieren oder den Zelltod auslösen. *Loss-of-function*-Mutationen inaktivieren diese Kontrolle und die Zellen teilen sich wieder schneller oder sterben nicht. Da diese Mutationen oft rezessiv sind, müssen beide Kopien des Gens betroffen sein. Wichtige Tumorsuppressorgene sind:

- RB: (*loss-of-function* führt zu Retinoblastom in der Netzhaut)
- P53: *guardian of the genome*, Transkriptionsfaktor, Zellzykluskontrolle, Apoptose, DNA-Reparatur
- PTEN: Protein- und Lipid-Phosphatase im PI3K-Akt-Signalweg
- Smad4: TGF-beta-Signalweg
- Tsc1 und Tsc2: Hemmung von mTOR
- APC: Kontrolle von Wnt

12.3 Klonale Selektion

Das Modell der klonalen Selektion besagt, dass ein Tumor aus einer Ansammlung von sogenannten *Driver*-Mutationen in der *most recent common ancestor-(MRCA)*-Zelle. *Driver*-Mutationen begünstigen oft das auftreten weiterer Mutationen, was zu immer neuen genetisch unterschiedlichen Zellpopulationen führt, die untereinander im Wettbewerb stehen können. Die kann bei Therapie zu Problemen führen, wenn sich die Zusammensetzung des Tumors zu therapieresistenten Zellen verschiebt. Die Entwicklung kann in Krebszellstammbäumen dargestellt werden und ist derjenigen in phylogenetischen Stammbäumen nicht unähnlich.



12.4 Hallmarks of Cancer

Nicht alle dieser Merkmale müssen in Krebs vorhanden sein und die Reihenfolge des Auftretens ist nicht entscheidend.

- *sustained proliferative signaling*: unabhängige Eigenversorgung mit wachstumsstimulierenden Signalen
- *evading growth suppressorss*: Umgehung von wachstumshemmenden Signalwegen
- *resisting cell death*: Resistenz gegen den programmierten Zelltod (Apoptose)
- *enabling replicative immortality*: Vermeidung der Seneszenz und des Zelltods durch extreme Telomer-Fehlfunktion
- *inducing angiogenesis*: Versorgung der Krebszellen mit ausreichend Nährstoffen aus der Blutbahn durch die Bildung von neuen Blutgefässen
- *invasion & metastasis*: Eindringen in andere Gewebe und Verteilung im Körper durch die Bildung von Metastasen
- *genome instability and mutation*: spezifische Mutationen und Instabilität des Genoms
- *tumor promoting inflamation*: dem Tumor nützliche Entzündungsprozesse in dessen Umgebung
- *deregulating cellular energetics*: Deregulation der zellulären Energieversorgung, z.B. Umstellung von aeroben auf anaeroben Metabolismus
- *avoid immune destruction*: Vermeidung des Immunsystems

12.5 The Cancer Genome Atlas

Das TCGA-Projekt hat sich zum Ziel gesetzt, die genetischen Veränderungen von 20 unterschiedlichen Tumortypen zu bestimmen. Dabei wurden je 500 Proben (10'000 insgesamt) auf Genom, epigenetische Markierungen, Anwesenheit nicht-kodierender RNA als auch RNA- und Proteinexpression untersucht. Eine der wichtigsten Erkentnisse des Projekts war, dass die Variation viel grösser als angenommen ist und rund 10 mal mehr Proben untersucht werden müssten, um die gesamte Diversität abzudecken.

12.6 Therapeutische Ansätze

Das zentrale Problem der Krebstherapie ist es, dass Krebszellen körpereigene Zellen sind.

12.6.1 Klassische Chemotherapie

Die verwendeten Wirkstoffe richten sich vorallem gegen schnell wachsende Zellen, also einen besonders aktiven Metabolismus haben. Die starken Nebenwirkungen wie Immunschwäche, Anämie, Haarausfall, Durchfall oder Übelkeit sind genau darauf zurückzuführen: Blutzellen, Haarfolikel und Zellen der Schleimhäute teilen sich ebenfalls schnell und werden somit auch Ziel des Wirkstoffes.

12.6.2 Zielgerichtete Krebstherapien

Mit der fortschreitenden Charakterisierung der verschiedenen Krebstypen wird es möglich, nur den Tumor anzugreifen. Ein Beispiel ist das Medikament PLX_{4032} (auch als Vemurafenib bekannt). Es wird seit 2011 gegen maligene Melanome eingesetzt und wirkt als selektiver Inhibitor des Onkogens B-Raf. Dieses kodiert für eine Ser/Thr-Kinase, die Zellzyklus und Wachstum reguliert. In Melanomen werden besonders oft V600E-Mutationen (Valin-Rest der 600. Aminosäure durch Glutamat ersetzt) in B-Raf beobachtet. Sie führen zur permanenten Aktivierung. PLX_{4032} bindet und inhibiert dadurch die mutierte Form von B-Raf. Der Wildtyp ist aber nicht beeinträchtigt. Bestimmte Darmkrebsvarianten, die auch die V600E-Mutation aufweisen, ist PLX_{4032} aber weitgehen ineffektiv. Wahrscheinliche Gründe sind, dass die Mutation dort nicht essentiel ist oder dass der Wirkstoff die Krebszellen nicht erreicht.
Ein weiteres Beispiel ist Erbitux, dass EGFT (*epidermal growth factor receptor*) inhibiert. Es ist sehr effektiv in der Bekämpfung von EGFT-exprimierenden Kolorektalkarzinomen ohne mutiertes K-Ras, nicht aber von vergleichbaren Lungenkrebsarten.
Zielgerichtete Krebstherapien sind sehr spezifisch, nicht frei von Nebenwirkungen und müssen oft von klassischer Chemotherapie begleitet werden. Die Entwicklung ist ausserdem sehr zeitaufwendig und kostspielig.

12.6.3 Personalisierte Krebstherapie

Dem Patienten werden Proben von gesundem Gewebe und des Tumors entnommen. Die daraus gezüchteten Zellkulturen werden dann auf die mehreren Hundert verfügbaren Medikamente *in vitro* getestet. Leider hat auch dieser Ansatz Probleme. Verschiedene Studien erhielten starkt unterschiedliche Resultate je nach dem wie das Medikament dosiert wurde.

13 Chemical Genetics

Small molecules sind Kohelnstoff-basierte niedermolekulare Verbindungen mit einem Gewicht unter 500 Dalton. Sie spielen wichtige Rollen in Rezeptorkontrolle, als sekundäre Messenger zwischen Proteinen, der Zellkommunikation, der Entwicklung und zur Kommunikation zwischen Organismen. Sie sind ausserdem vielversprechend als Medikamente. Allerdings können heutzutage erst rund 500 der 20'000 menschlichen Gene, die wiederum für mehr als 100'000 Proteine kodieren, mit *small molecules* angesteuert werden. *Chemical genetics* befasst sich mit der Untersuchung biologischer Systeme mithilfe von *small molecules*. Mit *forward chemical genetics* werden Funktionen von Genprodukten gesteuert, während bei *reverse chemical genetics* ein Protein mit einer *small molecules*-Bibliothek gescreent wird, um passende Liganden zu finde.

13.1 Vergleich mit anderen Methoden

Gegenüber genetischen Modifikationen hat *chemical genetics* den Vorteil, dass Untersuchungen *in vivo* durchgeführt werden können mit genauerer Kontrolle über Wirkungszeitraum und Dosierung. Durch den Einsatz von verschiedenen Liganden können ausserdem multifunktional Proteine untersucht werden.
Der grösste Nachteil ist der schmale Anwendungsbereich, der durch das limitierte Wissen über Protein-Bindungspartner beschränkt ist.

13.2 Anwendungen

Grosse Screens nach neuen Arzneimitteln sind mit hohen Kosten der *small molecule libraries* verbunden. Allerdings können sie auch weniger kostenintensiv durchgeführt werden und finden somit auch in der Forschung Anwendung zur Charakterisierung von Pathways.
Beispielsweise beschreibt Snyder et al. 2005 ein Screening nach *small molecules*, die Pigmentation in Albino-Melanozyten retten konnten. Dabei wurde Melanogenin als solches identifiziert und sein Zielprotein Prohibitin als bisher unbekannter Inducer von Pigmentation entdeckt.

13.3 small molecule Libraries

Um *small molecule libraries* aufzubauen, muss man zuerst einiges über die Struktur des Proteins und dessen natürlicher Liganden wissen. Bei unzureichender Kentniss, kann z.B. ein *high-throughput protein binding screen* durchgeführt werden. Zu Beginn wurden vorallem Peptide, die z.B. mit Merrifield's Peptid-Festphasensynthese hergestellt wurden, für die Bibliotheken verwendet. Nicht-peptidartige, organische *small molecules* haben aber den Vorteil durch die Plasmamembran wandern zu können. Deren Nachteil ist aber die relativ geringe Komplexität, was die spezifische Bindung an Proteine erschwert. Naturstoffe sind genug komplex, sind aber schwierig in genügend grossen Mengen zu erhalten. Ein guter Kompromiss sind Sammlungen von Naturprodukt-ähnlichen Stoffen. Diversitäts-orientierte Synthese ist eine weitere Methode, auch sie kommt aber mit einigen Schwierigkeiten.

13.4 Screening und Target Identification

In reversen Screens versucht man, ein *small molecule* zu identifizieren, welches spezifisch an ein Protein, DNA oder RNA bindet. Sie werden meist *in vitro* durchgeführt, um direkte Bindung und Dissoziation beobachten zu können. Eine Vorraussetzung sind aufgereinigte Proteine, DNA oder RNA. Basierend auf der Bibliotheksgrösse, den verwendeten Ressourcen und der Art des Targets wird dann eine Screeningmethode ausgesucht. Um die Bibliothek zu verkleinern, können virtuelle Screens am Computer (z.B. *pharmacophore-based virtual screening*) verwendet werden.
Forward Screens werden meist in Säugetier-Zelllinien oder ganzen Organismen durchgeführt. Dabei werden Moleküle auf erwartete Zellpermeabilität untersucht. Der *read-out* kann z.B. die Überlebensrate eines Organismus bzw. einer Zelllinie oder verändertes Signaling über einen bestimmten Pathway sein.

13.4.1 Small-Molecule Microarray

Ein Mikroarray wird mit rund 10'000 kovalent gebundenen *small molecules* bestückt, welche dann versuchen, ein bestimmtes Protein aus einer Lösung oder sogar einem Zelllysat zu binden. Anschliessend wird zuerst ein primärer und anschliessend ein Fluoreszenz-markierter, sekundärer Antikörper beigegeben und der Mikroarray mit einem Laser gescannt.

13.4.2 Cytoblot

Zellen werden auf einer 384-well Platte verteilt und mit unterschiedlichen *small molecules* inkubiert. Nach der Fixierung der Zellen werden sie mit primären und sekundären Antikörpern versehen. Letzterer ist dabei an *horseradish peroxidase* (HRP) und ein Reagenz für Chemoluminiszenz gekoppelt. Cytoblot ermöglicht es, Effekte auf die Zellphysiologie zu untersuchen, erlaubt aber keine intrazelluläre Lokalisation von Signalen oder Unterschiede zwischen Zellen zu finden.

13.4.3 Automatische Zell-Bildgebung

Technische Fortschritte in der automatisierten Mikroskopie erlauben es, einen Cytoblot-ähnlichen Screen automatisch per Bildgebungssoftware durchführen und auswerten zu lassen.

13.4.4 Target Identification

Der Vorteil von Forward Chemical Genetics ist die Möglichkeit, für fast jeden Phänotyp einen passenden Screen zu entwickeln. Die Zielidentifikation ist dabei aber eine grosse Herausforderung. Bei Affinitäts-basierten Methoden wird das *small molecule* als erstes immobilisiert, z.B. auf Glasträgern, magnetischen Beads, Affinitätschromatographie-Beads oder an Streptavidin-Beads mithilfe von Biotinmarkern. Meist mithilfe von Affinitätschromatographie wird dann ein Bindungspartner aus einem Zelllysat oder Proteinextrakt gesucht. Die am stärksten bindenden Proteine werden mit SDS-PAGE separiert und mit Massenspektrometrie bestimmt. Nachteile sind die Identifikation von Proteinen mit geringer Affinität und die Sensitivität von MS *read outs*.
Nach der Auswahl bestimmter Moleküle werden *in vitro* sogenannte primäre Verfahren, am häufigsten quantifizierbare Fluoreszenz-Read-outs, angewandt. Eine Beliebte Methode ist die Fluoreszenz-Polarisation, bei der ein Fluorophor-markierter Inhibitor von Protein-Proteininteraktionen je nach Zustand polarisiertes Licht reflektiert oder nicht. Ein anderes Verfahren ist der Förster-Resonanz Energietransfer (FRET). Dabei wird ein Donor-Fluorophor mit einer bestimmten Wellenlänge angeregt. Dessen Fluoreszenz regt dann ein Akzeptor-Fluorophor an. So kann man die Entfernung zwischen zwei Molekülen messen und z.B. Heterodimerstabilisierende *small molecules* identifizieren.
Sekundäre Verfahren sind dazu da, primäre Verfahren zu verifizieren. Dies kann z.B. mit biophysikalischen Messungen der Bindungskonstanten und thermodynamischer Parameter erreicht werden. Beispiele sind die Kernspinresonanz-Spektroskopie oder die isotherme Titrationskalorimetrie, bei der die Änderung der Wärme bei Bindung gemessen wird.

13.5 Metabolic Mining

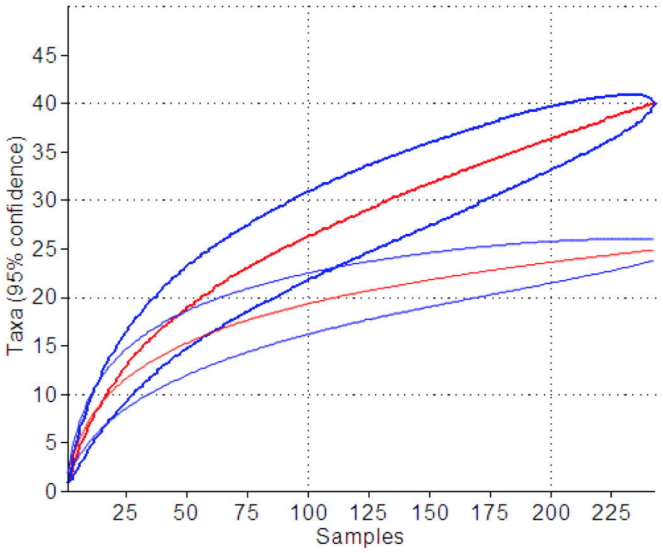
Small molecules von Mikroorganismen wie Peptide oder Polyketide haben oft grosses Potential für die Entwicklung neuer Arzneistoffe. Die dafür kodierenden Gene liegen oft in Clustern vor und die entsprechenden Enzyme sind modular organisiert. *Top-down* Ansätze beginnen beim Organismus und versuchen diesen dazu zu stimulieren, Naturstoffe zu produzieren. Die biologsichen Proben werden kultiviert oder direkt auf spezifische Bioaktivität untersucht. Vielversprechende Moleküle werden dann strukturell charakterisiert.
Top-down Ansätze funktionieren am besten bei kontinuierlich produzierten Stoffen. Dabei werden eine ganze Reihe von verschiedenen Sampling- und Kultivierungsmethoden verwendet, um Stresssituationen zu simulieren.
Genom-basierte *bottom-up* Ansätze nutzen die Tatsache aus, dass für Naturstoffe verantwortliche Gene in Clustern vorliegen. Mit rechengestützen *in silico*-Methoden (am Computer) identifizierte Cluster werden mit Genmanipulationsmethoden zur Transkription, Translation und Synthese des Stoffes gebracht. Da diese Cluster oft inaktiv sind unter Laborbedingungen, funktionieren Knockouts nicht und es werden stattdessen Knockins verwendet. Ein Promotor aktiviert dabei das Cluster. Anti-SMASH (*antibiotics & secondary metabolite analysis shell*) ist eine Software, die alle bekannten Klassen biosynthetischer Cluster erkennen, sie funktionell annotieren und die chemische Struktur vorhersagen kann. Ausserdem kann mithilfe eines eingebauten Gencluster-Analysetools auf Verwandschaft von Clustern und damit Funktionen von Genen geschlossen werden

14 Metagenomik

Ein Gramm Waldboden enthält 4×10^7 prokaryotische Zellen und 2000-18′000 verschiedene Genome. Ein Milliliter Meerwasser der Sargassosee beherbergt rund eine Million Baktarien und hat eine durchschnittliche Genomgrösse von zwei Millionen Basenpaaren. Diese Gesamtheit der genomischen Information aller Mikroorganismen einer bestimmten Lebensgemeinschaft wird als Metagenom bezeichnet. Die Metagenomik beschäftigt sich mit dessen Analyse. Mit Shotgun-Sequenzanalysen konnte die enorme funktionelle Gendiversität in Mikroorganismen gezeigt werden. Obwohl viele der Fragmente bei einem Shotgun-Ansatz nicht einer bestimmten Spezies zugeordnet werden können, ermöglicht die Metagenomik Schlüsse über potentiell neue Biomoleküle zu ziehen, genomische Beziehungen zwischen Funktion und Phylogenie und evolutionäre Profile von Struktur und Funktion einer mikrobiellen Lebensgemeinschaft aufzudecken.

14.1 Diversität und Rarefaction-Analyse

Zur Bestimmung der Diversität einer Umweltprobe wird oft ribosomale RNA (rRNA) analysiert. Durch ihre essentielle und konservierte Funktion unterliegt sie kaum horizontalem Gentransfer und ist somit als molekularer Marker gut geeignet. Kleine rRNAs (5S und 5.8S) liefern zu wenig phylogenetische Informationen, während grosse rRNAs (23S und 28S) schwieriger zu untersuchen sind. Die Untersuchung von 16S rRNA bei Prokaryoten bzw. 18S rRNA bei Eukaryoten hat sich als Methode durchgesetzt. Die Rarefaction-Analyse hat sich als Hilfsmittel zur Bestimmung des Anteils einer Probe am gesamten Artenreichtum etabliert. Dazu wird die Anzahl Spezies als Funktion der Anzahl Proben geplottet. Die Kurven steigen meist zu beginn stark an und flachen danach ab. Je kleiner die Steigung, desto weniger trägt das Sample zur Identifikation von *operational taxonomic units* (OTUs) bei. Mit dieser Methode kann man die genetische Diversität zwischen Studien vergleichen, da gut auf die Gesamtzahl der OTUs extrapoliert werden kann.



14.2 Metagenomische Libraries

Die Konstruktion von Libraries erlaubt die Analyse ohne die Mikroorganismen vorher kultivieren zu müssen. Für die Suche nach neuen Biomolekülen kommen entweder funktions- oder sequenzbasierte Screens zum Einsatz.

14.2.1 Extraktion und Aufbereitung metagenomischer DNA

Extraktion von DNA aus Wasserproben erfordert oft grosse Wassermengen, während Bodenproben anorganische Stoffe wie Huminsäuren oder DNAsen enthalten, die den Prozess erschweren. Mit mehreren aufeinander folgenden Extraktionsmethoden wie Gelelektrophorese kann die DNA entkontaminiert werden, es geht aber auch immer etwas genetisches Material verloren. Auch schwierig ist die Extraktion aus robusten Organismen, wie eingekapselten Bakterien oder Sporen.

14.2.2 Klonierungsvektoren

Je nach Grösse der zu klonenden DNA-Fragemente werden Plasmide (<15kb), Cosmide, Fosmide (beide <40kb) oder künstliche Bakterienchromosomen (BACs, >40kb) als Vektoren verwendet. Während etwas 20-70 Cosmide pro Zelle vorliegen, existiert normalerweise nur ein Fosmid in einer Zelle. Fosmide sind etwas stabiler und durch Arabinose-Zugabe kann deren Anzahl erhöht werden. Die Grösse des Vektorsystems hängt von DNA-Qualität, Zielgenen und Screening-Strategie ab. Libraries mit kleinen DNA-Inserts können der Identifikation von Biomolekülen dienen. Libraries aus DNA-Inserts in Plasmiden erfordern aber auch 3-20x mehr Klone. Solche mit grossen DNA-Inserts dagegen sind nützlich zur Identifikation von Biosynthese-Pathways. Dabei ist praktisch, dass bei Bakterien funktional verknüpfte Gene oft in Operons liegen.

14.2.3 Transformation eines Hosts

Der am häufigsten verwendete Host ist *E. coli*. Mutationen die Rekombination (*recA*) und DNA-Degradation (*endA*) hemmen als auch solche, die ein rekombinantes blau/weiss Screening (*lacZ*) erlauben, sind wünschenswert. Metagenomische DNA wird meist per Elektroporation eingefügt. Eine weitere Methode ist es Phagen zu verwenden. Die Expression von Genen aus weiter entfernt verwandten Organismen stellt aber ein Problem dar. Dafür werden dann andere Hosts verwendet.

14.2.4 Sequenzbasierte Metagenomik

Bei der sequenzbasierten Screens werden Schlussfolgerungen auf Funktion und Verwandschaft aus einer Sequenzanalyse gezogen. Mit Primern für konservierte Abschnitte von bereits bekannten Genen, lassen sich neue Varianten finden. Grosse Libraries können auch ohne Vorselektion von Klonen sequenziert werden. Dabei werden Primer für Sequenzen, welche die geklonte flankieren, verwendet.

14.2.5 Funktionsbasierte Metagenomik

Funktionelle Metagenomik identifiziert Klone mit einer bestimmten durch metagenomische DNA hervorgerufenen Aktivität. Man benötigt im Gegensatz zu sequenzbasierten Methoden keine bekannten Homologien. Sie ist der vielversprechendste Ansatz zur Entdeckung neuartiger Biomoleküle. Um bestimmte Molekülklassen zu analysieren, können Gemeinschaften vor der Konstruktion der Library manipuliert werden.

14.3 Datananalyse

14.3.1 Assemblierung

Im Gegensatz zur relativ einfachen Assemblierung einzelner Genome ist die vollständige Rekonstruktion eines Metagenoms meist nicht möglich. Deswegen können meist nur Sequenzen in den Bereichen 25-75bp (SNPs, short frameshift mutations), 100-400bp (kurze funktionale Elemente), 500-1'000bp (Domänen, single domain genes) und höchstens 1'000-5'000bp (kurze Operone, multidomain genes) identifiziert werden.
Die Abdeckung C mit Read Length L, Anzahl Reads N und Genomgrösse G bezeichnet die durchschnittliche Häufigkeit, mit der ein einzelnes Nukleotid sequenziert werden muss, um das komplette Genom zu identifizieren. Die prozentuale Abdeckung eines Genoms ist P_0. Die benötigte Anzahl Reads für eine bestimmt Abdeckung ist also N. Dabei ist die Metagenomgrösse G_m bzw. $\hat{G}_m$ mit der jeweiligen Häufigkeit der Spezies p_i ($\sum_{i=1}^{l} p_i = 1$). Mit Spezies-spezifischen rDNAs können die p_i-Werte abgeschätzt werden, dennoch sind adäquate Abdeckungen für artenreiche Gemeinschaften schwierig zu erreichen.

$$C = \frac{L * N}{G} \qquad P_0 = 1 - e^{-C}$$

$$N = -\frac{ln(1 - P_0) * G}{L} \qquad G_m = \sum_{i=1}^{l} n_i G_i$$

$$\hat{G}_m = p_1 G_1 + p_2 G_2 + \ldots + p_l G_l$$

Bei der Assemblierung besteht die Gefahr, dass Sequenzen verschiedener OTUs zu Chimären zusammengefügt werden. Für die Assemblierung einzelner, Sanger-sequenzierter Genome entwickelte Algorithmen wie Phrap, Forge, Arachne, JAZZ oder Celera Assembler machen einen relativ guten Job mit metagenomischen Fragmenten.

14.3.2 Binning

Die Zuordnung von Sequenzierdaten zu den entsprechenden OTUs wird als Binning bezeichnet. Phylogenetische Markergene sind oft aber nicht vorhanden, da sie nicht assembliert wurden oder sogar im gesamten Datensatz fehlen. Dabei gibt es folgende Methoden:

- **Taxonomie-abhängige Methoden** vergleichen Reads mit Referenzdatenbanken. Bei zu geringer Ähnlichkeit werden Sequenzen als *unassigned* klassifiziert.
  - **Alignment-basierte Methoden** verwenden BLAST um einzelne Reads mit Referenzsequenzen (bei NCBI, EMBL oder UniProt öffentlich vorhanden) zu alignen. Danach werden Reads mit verschiedenen Hit-Sequenzen aus der Datenbank in taxonomische Gruppen eingeordnet. Da die Hits oft nicht gut genug sind, wird die LCA-Methode (*lowest common ancestor*) angewandt. Dabei werden Reads dem Vorfahren des besten Hits zugeordnet.
  - **Kompositions-basierte Methoden** vergleichen mit Datenbanken über Eigenschaften wie GC-Gehalt, Codon-Verwendung und Oligonukleotidmuster gemacht. Die Methode ist schneller als Alignment-basierte Techniken, braucht aber längere Reads.
  - **Hybridmethoden** sind Kombinationen aus den beiden anderen Methoden. Der SPHINX-Algorithmus führt ein sogenanntes Zweiphasen-Binning durch. Dabei wird zuerst die Anzahl Referenzsequenzen verringert. Erst im zweiten Schritt folgt das Alignment.
- **Taxonomie-unabhängige Methoden** verwenden intrinsische Informationen statt Referenzdatenbanken. Der TETRA-Algorithmus errechnet paarweise Korrelationsmuster verschiedener Tetranukleotide für die einzelnen Reads, da diese optimales Unterscheidungspotential aufweisen. Die Reads werden anschliessend in Gruppen geordnet.

14.3.3 Annotation

Die erste Phase der Annotation besteht in der Zuweisung von biologischen Funktionen zu einzelnen ORFs. Bei metagenomischen Proben stellt die häufige Unvolständigkeit von ORFs und grosse Teile ohne Homologe ein Problem dar. Im zweiten Schritt werden Gene identifiziert, die ein biologisches Netzwerk bilden. Auch hier liegen die Probleme in der Schwierigkeit ORFs einer Spezies zuzuordnen. Mit *six-frame* Translationen können ORFs auf genug langen Reads und möglicherweise partielle ORFs auf kürzeren Reads gefunden werden. Nicht-assemblierte Reads können mit BLAST-Hits Rückschlüsse erlauben.
MG-RAST (*metagenomic rapid annotations using subsystems technology*) ist eine open source Webapplikation, die Sequenzvergleiche auf Nukleotid- und Proteinlevel macht.
RAMMCAP (*rapid analysis of multiple metagenomes with a clustering and annotation pipeline*) verwendet den CD-HIT-Algorithmus um translatierte ORFs nach Sequenzähnlichkeit zu kategorisieren. Je grösser die Cluster, desto wahrscheinlicher befindet sich dort ein wahrer ORF. So werden nur ein Teil der Fragmente mit der grössten Ähnlichkeit mit Datenbanken verglichen.