Forschungstechniken der Genomik

Einführung

Genomik ist die Anwendung von biologischen Forschungsansätzen und Labormethoden auf der Ebene des **gesamten** Genoms. In der Genomik wird also z.B. nicht mehr die Expression eines einzelnen Gens gemessen oder der Effekt der Ausschaltung eines einzelnen Gens untersucht, sondern man misst die Expression **aller** 10-20'000 Gene eines Organismus oder man schaltet systematisch **jedes** einzelne Gen eines Organismus aus und untersucht den Effekt **aller** resultierenden Mutanten.

Dieser genomische Ansatz ist noch relativ jung, findet aber ausgehend von der Genetik Einzug in alle anderen Bereiche der biologischen Forschung (Proteomik, Metabolomik, Evolutionsforschung, Strukturforschung etc.).

In den verschiedenen Abschnitten des Kurses wird das Thema dieses Übergangs von Einzelgen- zu genomweiten Ansätzen immer wieder auftauchen und die daraus resultierenden neuen Möglichkeiten für die biologische Forschung werden dort jeweils aufgegriffen.

In dieser Lektion betrachten wir nun einige der Schlüsseltechnologien die diese genomweiten Ansätzen möglich gemacht haben. Beim Lesen dieses Handouts ist es empfehlenswert das Augenmerk auf die grundlegenden Konzepte hinter den besprochenen Methoden und weniger auf deren genaue technische Details zu richten. Dafür gibt es zwei Gründe. Zum einen führen die hohen Gerätekosten und technischen Ansprüche dazu, dass diese Methoden (häufig auch als Assays bezeichnet) nur selten von den jeweiligen Forschungslabors selber durchgeführt werden. Stattdessen an zentralisierte Analyselabors (sogenannten *core facilities,* z.B. dem Functional Genomics Center Zurich) bzw. an spezialisierten Firmen ausgelagert werden. Dort kümmern sich Spezialisten um diese Details. Zum anderen unterliegen diese technischen Details einem extrem rapiden Wandel. Bestimmte grundlegende Konzepte (e.g. Kodierung der Identität eines Moleküls durch dessen xy-Postion auf einer festen Oberfläche, oder die Verknüpfung von lesbarer Information und physischer Struktur bei DNA) haben hingegen über mehrere technische Generationen nichts an Erklärungskraft eingebüsst.

Genotypisierung

Ein zentrales Ziel der Genetik ist es, zu verstehen wie die von Generation zu Generation vererbte genetische Information (der Genotyp) den Aufbau und das Verhalten (den Phänotyp) eines Organismus beeinflussen. Seit Mitte des letzten Jahrhunderts ist bekannt, dass die erbliche Information in der linearen Sequenz der DNA gespeichert ist. Inzwischen ist es auch technisch möglich, die gesamte DNA Sequenz eines Organismus zu bestimmen und so jeden Aspekt des Genotyps in diesem Organismus zu bestimmen. Die Sequenziertechniken, die dies ermöglichen, haben Sie bereits im letzten Semester im Kurs „Methoden der Biologischen Analytik“ kennen gelernt.

Trotz der immensen technischen Fortschritte in diesem Bereich ist die komplette Sequenzierung des gesamten Genoms aller Individuen in einer Studie dennoch recht langwierig, aufwendig und teurer. Dabei ist die eigentliche Sequenzierung, also die Generierung der primären Rohdaten (also der sogenannten *reads*) fast noch das kleinste Problem. Die Speicherung, Verwaltung, Analyse und Interpretation von genomweiten Sequenzdatensätzen stellt hingegen weiterhin eine erhebliche Herausforderung dar.

Aus diesen Gründen kommen in vielen genetischen Studien weiterhin Genotypisierungsmethoden zur Verwendung in denen der Genotyp der untersuchten Individuen nur an bestimmten strategisch ausgewählten Stellen im Genom gemessen wird. Zwei besonders populäre Arten dieser Genotypisierungstechniken sind in den nachfolgenden Abschnitten besprochen.

PCR-basierte Genotypisierung

PCR-basierte Genotypisierungstechniken sind darauf ausgelegt, einzelne genetische Variationen an einer einzigen Stelle im Genom zu untersuchen. Ein Beispiel hierfür wäre die Untersuchung eines genetisch bestimmten Stoffwechseldefekts für den die ursächliche genetische Variation (häufig hervorgerufen durch einen SNP) bereits genau bekannt ist. In solchen Fällen geht es bei der Genotypisierung also nur darum herauszufinden, welche von zwei möglichen Allelen (Variante A oder B) an genau dieser Stelle des Genoms vorliegt. Hierzu führt man zwei nahezu identische PCR Reaktionen durch. Der einzige Unterschied zwischen den beiden Reaktionen ist die Sequenz einer der beiden verwendeten Primer. Der zweite Primer ist in beiden Reaktionen identisch. In der ersten Reaktion (Reaktion A) verwendet man einen Primer der perfekt komplementär zu Variante A des untersuchten Gens ist und in der zweiten Reaktion (Reaktion B) einen Primer der perfekt komplementär zu Variante B ist. Als Template für die PCR Reaktion verwendet man in beiden Fällen die genomische DNA des zu untersuchenden Individuums. Trägt das Individuum zwei Kopien der Genvariante A in seinem Genom, erwarten wir, dass der Primer in Reaktion A perfekt an seine Targetsequenz im Genom

bindet und die PCR Reaktion deshalb effizient verlaufen wird. In Reaktion B entsteht beim Anbinden des Primers an seine Targetsequenz ein Missmatch. Die Bindung des Primers an seine Targetsequenz fällt also weniger stark aus und die Produktion des PCR Produkts wird daher auch weniger effizient sein. Trennt man dann die Produkte der beiden PCR Reaktionen durch Gelelektrophorese auf, erwarten man für Reaktion A eine klar sichtbare Bande des PCR Produkts. Die äquivalente Bande in Reaktion B sollte aufgrund der weniger effizienten Bindung des Primers hingegen kaum oder gar nicht sichtbar sein.

Würde man dieselben zwei Reaktionen mit der genomischen DNA eines Individuums durchführen, welches sowohl vom Vater als auch von der Mutter Genvariante B geerbt hat, so würde man hingegen eine klar sichtbare Bande für Reaktion B aber keine Bande für Reaktion A erwarten.

Frage: Welches Bandenmuster würde man erwarten, wenn das untersuchte Individuum an dem untersuchten Locus heterozygot ist?

Der oben beschriebene Test veranschaulicht recht klar das fundamentale Prinzip eines PCR-basierten Genotypisierungstests. Diese Variante einer PCR-basierten Genotypisierung wird in dieser Form aber nur noch relativ selten angewendet. Die Gründe hierfür sind, dass die PCR-Reaktionen für diese Art von Tests zunächst relativ aufwendig optimiert werden müssen, pro Probe zwei separate PCR Reaktionen ausgeführt werden müssen und der Nachweis der PCR-Produkte einen separaten Gelelektrophoreseschritt benötigt. Für grössere Studien wo Effizienz und Reproduzierbarkeit eine besonders wichtige Rolle spielen, werden daher typischerweise andere Varianten von PCR-basierten Genotypisierung verwendet. Diese Varianten sind, vom Konzept her, etwas komplizierter, sind in der Anwendung aber deutlich effizienter und zuverlässiger.

Hierbei ist vor allem die fluoreszenz-basierte TaqMan Methode (Abbildung 1) beliebt. Diese TaqMan Methode basiert auf einer PCR Reaktion in der vier Primer verwendet werden. Zwei dieser Primer sind konventionelle PCR-Primer welche den zu genotypisierenden Locus flankieren. Die beiden anderen Primer (Primer A und Primer B) sind jeweils an einem Ende mit einem Fluorophore (z.B. Primer A grüner Fluorophore und Primer B roter Fluorophore) und am anderen Ende mit einem Quencher-Molekül gelabelt. Dadurch, dass die Fluorophore und die Quencher über den Primer aneinandergekoppelt, also räumlich nahe bei einander sind, unterdrückt der Quencher die Fluoreszenz des Fluorophors. Diese gelabelten Primer sind von der Sequenz her so gewählt, dass sie direkt an den zu genotypisierenden Locus binden. Dabei bildet Primer A einen perfekten Match mit der Sequenz der Genvariante A, erzeugt aber mit Genvariante B einen Einzelbasenmissmatch. Für Primer B ist die Situation genau umgekehrt

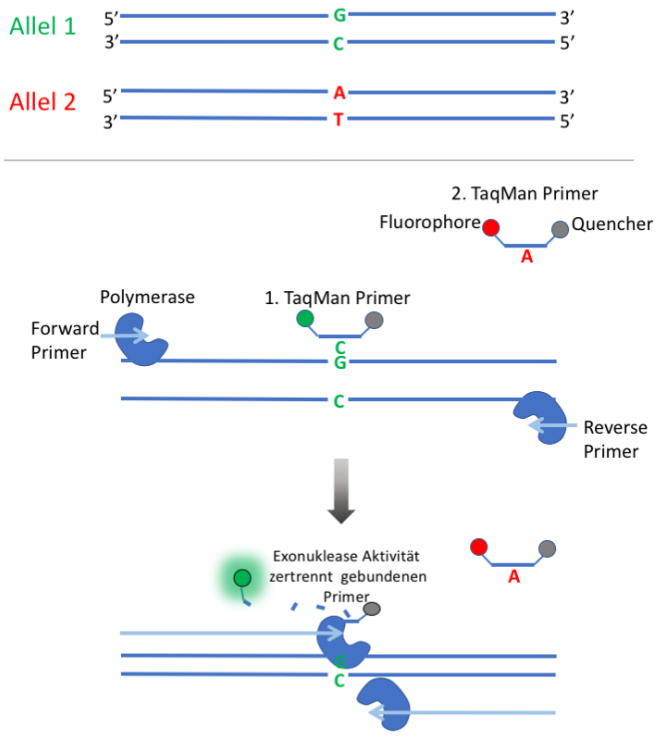

Abbildung 1 *TaqMan eine PCR-basierte Methode zur Genotypisierung. Das obere Panel zeigt die zwei möglichen Allele der untersuchten DNA Sequenz. Die TaqMan PCR Methode (unteres Panel) verwendet neben den üblichen Komponenten einer PCR Reaktion (Template, Forward und Reverse Primer etc.) zusätzlich zwei TaqMan Primer. Diese Primer sind jeweils an einem Ende mit einem Fluorophore und am anderen Ende mit einem Fluoreszensquencher verknüpft. TaqMan Primer 1 ist komplementär zu Allel 1 und Primer 2 zu Allel 2. Während der PCR Reaktion wird die zwischen den Primern liegende Templatesequenz amplifiziert und entsprechen dessen Sequenz bindet sich entweder TaqMan Primer 1 oder 2 (in hier gezeigten Fall Primer 1). Im nächsten PCR Zyklus zertrennt die Exonuklease Aktivität der Polymerase den gebundenen Taqman Primer in einzelne Nucleotide. Dadurch trennen sich Fluorophore und Quencher wodurch der Fluorophore aktiv wird. Der nicht gebundene Primer bleibt intakt und fluoresziert nicht. Die Farbe der in dieser Reaktion entstehenden Fluoreszenz zeigt also das Allel der Template DNA an.*

(perfekter Match mit Genvariante B aber Einzelbasenmissmatch mit Genvariante A). Führt man nun eine PCR-Reaktion durch, in der Template DNA eingesetzt wird, die ausschliesslich Genvariante A enthält, so werden durch die PCR Reaktion der beiden äusseren Primer PCR Produkte mit der Sequenz der Genvariante A erzeugt. Der gelabelte Primer A bindet aufgrund der perfekten Sequenzkomplementarität besonders stark an diese PCR

Produkte. Im nächsten PCR Zyklus trifft die DNA-Polymerase während der Synthese des neuen Strangs nun auf diesen an das Template gebundenen gelabelten Primer und zertrennt diesen mittels seiner natürliche Exonuclease Aktivität in einzelne Nukleotide. Diese einzelnen Nukleotide diffundieren nun unabhängig voneinander durch die Reaktionsmischung. Der Abstand zwischen den beiden nimmt zu und der Quencher kann die Fluoreszenz des Fluorophores nicht mehr unterdrücken. Bei entsprechender Beleuchtung entsteht nun ein für Gen Variante A spezifisches (in diesem Fall grünes) Fluoreszenz Signal. Würde die Template-DNA hingegen nur Genvariante B enthalten, so würde bevorzugt Primer B an die entstehenden PCR Produkte binden, abgebaut werden und so ein rotes Fluoreszenzsignal erzeugen. Mit speziellen PCR Maschinen (sogenannten *real-time* PCR Geräten) kann man die Intensität sowohl des grünen als auch des roten Fluoreszenzsignals in derselben Reaktion „live" während der PCR Reaktion mitverfolgen. Der Genotyp des getesteten Individuums ergibt sich dann aus der relativen Intensität des roten und grünen Fluoreszenzsignals. Diese TaqMan Methode bietet eine Reihe von Vorteilen. Zum einen entfällt der zeitaufwendige Gelelektrophorese Schritt, was die Geschwindigkeit und Effizienz des Assays stark verbessert. Zum anderen finden die Reaktion zur Quantifizierung von Genvariante A und B in ein und demselben Reaktionsvolumen, also unter identischen Bedingungen, statt. Eventuelle Fluktuation in der Temperatur oder Pipettierfehler während der Herstellung der Reaktionsmischung haben daher den gleichen Einfluss auf beide Reaktion, gleichen sich also gegenseitig aus.

Frage: Beide der hier vorgestellten PCR-basierten Genotypisierungsmethoden verwenden allel-spezifische Primer. Was passiert im Laufe der PCR-Reaktionen mit diesen allel-spezifischen Primern? Wo liegt der Unterschied zwischen den beiden Methoden?

Microarray-basierte Genotypisierung

Moderne microarray-basierte Genotypisierungsmethoden bestimmen den Genotyp an ca. 1 Million über das Genom verteilter Loci. Sie erreichen dies durch die parallele Durchführung von ebenso vielen, stark miniaturisierten Genotypisierungsreaktionen auf einem einzigen Chip (oft als SNP-Chip bezeichnet). Microarray-basierte Genotypisierungsmethoden decken also zwar das gesamte Genom ab, bestimmen aber (anders als die Sequenzierung) den Genotyp nicht an jeder einzelnen Base des Genoms sondern nur an strategisch ausgewählten Loci. In diesem Sinne nehmen Microarrays unter den Genotypisierungsmethoden eine Zwischenposition zwischen den PCR-basierten und den *whole-genome* Sequenzierungsmethoden ein.

Frage: An jeder wievielten Base des menschlichen Genoms bestimmt ein solcher SNP-Chip im Durchschnitt den Genotyp?

Ursprünglich wurden die in diesen Experimenten verwendeten Microarrays noch von den Forschungslabors selbst hergestellt. Inzwischen sind die kommerziell angebotenen Genotypisierungsmicroarrays aber so leistungsfähig und günstig geworden, dass selbst hergestellte Microarrays nur noch extrem selten verwendet werden.

Die Details der eigentlichen Genotypisierungsreaktion unterscheiden sich von Hersteller zu Hersteller und unterliegen ständiger Optimierung und Anpassung. Viele der zugrundeliegenden Prinzipien sind aber allen diesen Varianten gemein. Stellvertretend für alle anderen Varianten wird hier die von der Firma Illumina entwickelte besonders populäre Infinium II Methode (Abbildung 2) besprochen.

Die Basistechnologie von Microarrays besteht in der Immobilisierung von Clustern von relativ kurzen (ca. 50 Basen langen) einzelsträngigen DNA Molekülen (sogenannten Oligos) auf einer Oberfläche. Der Trick besteht darin, dass innerhalb eines Clusters alle Oligos dieselbe Sequenz besitzen, dass die Sequenzen aber von Cluster zu Cluster unterschiedlich sind. In einem Genotypisierungs Microarray sind die Oligo Sequenzen der Cluster jeweils komplementär zu der DNA Sequenz von unterschiedlichen über das ganze Genom verteilten Loci.

Zertrennt man nun die genomische DNA der zu untersuchenden Probe in kurze Fragmente, denaturiert diese Fragmente und gibt diese dann auf das Array, so werden die genomischen DNA Fragmente durch spezifische Watson-Crick Basenpaarung selektiv an die zu ihnen komplementären Oligos gebunden. Auf diese Weise erhalten wir Cluster von doppelsträngigen DNA Segmenten die aus genomischen DNA Fragmenten und den zu ihnen komplementären Oligos bestehen.

Zur Genotypisierung von SNPs wählt man die Sequenz des immobilisierten Oligos jeweils so, dass das 5' Ende des Oligos genau mit demjenigen Nukleotid der genomischen DNA Sequenz gepaart ist, welches unmittelbar vor dem Nukleotid liegt, dessen Identität wir bestimmen wollen. Man benutzt nun eine Polymerase und speziell präparierte Nucleotide um eine einzelne Base an das 5'-Ende des Oligo's anzuhängen. Die Polymerase baut dabei dasjenige Nucleotid ein, welches komplementär zum Nucleotid auf der genomischen DNA Fragment ist. Da die Nucleotide mit spezifischen Labeln versehen sind, kann man anschliessend (analog zur *sequencing-by-synthesis* DNA Sequenziermethode) mittels optischem *Read Out* feststellen, welche Base an dieser, durch die Oligosequenz spezifizierten Stelle des Genoms vorgelegen hat. War das untersuchte Individuum heterozygot an diesem Locus, so wird das optische Signal eine 50/50 Mix der Signale für die beiden möglichen Varianten sein.

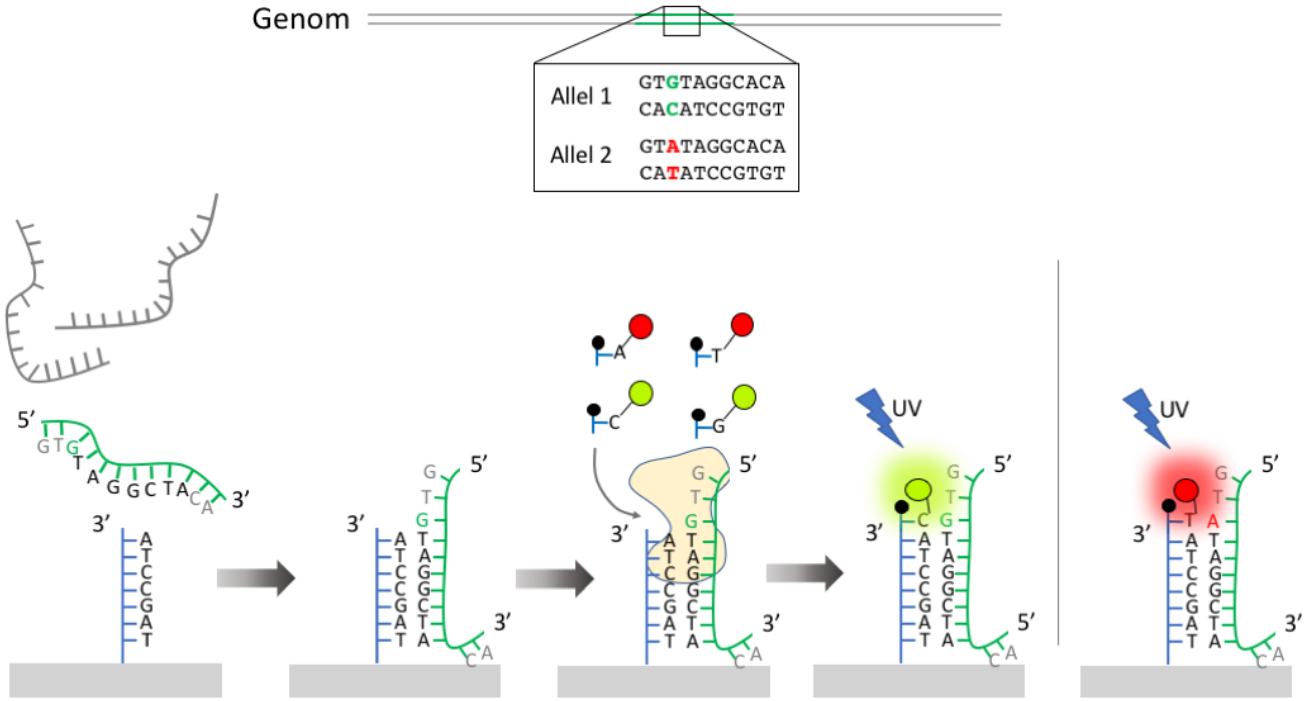

Abbildung 2 *Microarray-basierte Genotypisierung. Am oberen Abbildungsrand ist ein Ausschnitt aus einem Genom gezeigt. Von dem in grün markierten Gen gibt es zwei Allele. Die Abbildung erklärt wie das Illumina Infinium II Assay bestimmt, welches der beiden Allele in einer Probe vorliegt. Die genomische DNA wird in kurze Fragmente zertrennt, denaturiert und auf Microarray gegeben. Das aus dem Genom stammende DNA Fragment (grün) des zu untersuchenden Gens bindet spezifisch an das entsprechende an der Arrayoberfläche immobilisierte Oligo (blau). Die Sequenz des Oligos ist dabei so gewählt, dass die zu genotypisierende Base ungepaart bleibt. Wird nun eine Polymerase (hellgelb) und mit einem Marker versehene Nukleotide hinzugegeben, baut die Polymerase dasjenige Nukleotid in die Oligo sequenz ein, welches komplementär zu der genotypisierenden Base ist. Ein Blocker (kleiner schwarzer Kreis) verhindert die Einfügung weiterer Nukleotide. Die Farbe des an das eingefügte Nukleotid gebundenen Markes zeigt nun die Identität der zu genotypisierenden Base an. Das rechte Panel zeigt das Resultat des Assays wenn die Probe das Allel2 enthalten hätte. Auf einem modernen Genotypisierungsmicroarray können mehr als eine Million dieser Assays parallel zu einander durchgeführt werden.*

Der Kostenvorteil von microarray-basierter Genotypisierung gegenüber der vollständigen Sequenzierung des gesamten Genoms ist in den letzten Jahren kontinuierlich geschrumpft. Dennoch erfreut sich microarray-basierte Genotypisierung weither hin grosser Beliebtheit, vor allem bei klinischen Anwendungen. Diese Beliebtheit erklärt sich vor allem daraus, wie einfach, robust und reproduzierbar die Interpretation von SNP-Chip Daten im Vergleich zur Interpretation von *whole-genome* Sequenzierdaten ist.

Frage: Ist es mit der hier beschriebenen microarray-basierten Genotypisierungsmethode möglich neue, zuvor unbekannte Single Nucleotide Polymorphismen zu entdecken? Warum?

Getargete DNA Sequenzierung (z.B. Exom Sequenzierung)

Wie schon öfter angesprochen ist die Sequenzierung und vor allem die bioinformatische Aufarbeitung von kompletten Säugetiergenomen weiterhin recht zeit- und kostenaufwendig. DNA-Sequenzierung ist aber die einzige Möglichkeit neue, vor der Analyse unbekannte, genetische Variation aufzuspüren. In vielen genetischen Studien ist aber genau diese Aufspürung von *de novo* Mutationen gefordert. Diejenigen *de novo* Mutation mit den stärksten phänotypischen Effekten fallen dabei häufig in die proteinkodierenden Regionen des Genoms (z.B. neu entstandene Stop-Codons welche die Funktion eines Proteins komplett zerstören). Zudem sind solche, eine protein-kodierende Sequenz verändernde Mutationen in der Regel recht einfach zu interpretieren. In vielen Fällen, z.B. bei der Suche nach Mutationen, die schwerwiegende aber sehr seltene genetische Krankheiten auslösen, würde es daher durchaus Sinn machen, zunächst nur das Exom zu sequenzieren, also denjenigen ca. 1,5% des Genoms, die direkt Proteine codierenden.

Eine solche getargetete DNA Sequenzierung kann dadurch erreicht werden, dass man eine genomische DNA Library vor der Sequenzierung einem Selektionsschritt unterzieht. Hierzu wird die genomische DNA, wie

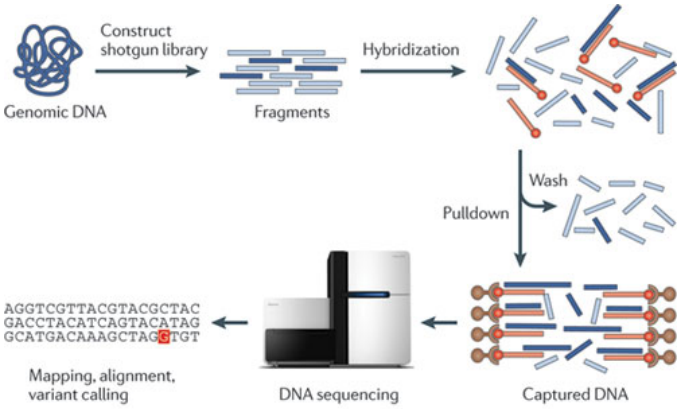

Abbildung 3 *Für die Exome Sequenzierung wird die DNA Fragment Library einem Selektionsprozess unterzogen, der gezielt protein-kodierende DNA Fragmente (dunkel blau) anreichert. Dazu mischt man die Fragment Library mit einem Pool von Oligos (orange) deren Sequenzen komplementär zu den protein-kodierenden Regionen des Genoms sind. Diese oligos sind an einem Ende mit Biotin gelabelt. Die so entstandenen Hypride können somit an Biotin-dekorierte Beads (braun) gebunden und isoliert werden. Die so erhaltenen Fragmente von protein-codierender genomischer DNA können nun sequenziert werden. (Bildquelle Bamshad et al. 2011 Nature Reviews Genetics)*

gewohnt in einige hundert Basen lange Stücke fragmentiert und denaturiert (also in Einzelstränge zerlegt). Diesem Mix aus einzelsträngigen DNA Fragmenten werden nun Pools von Oligos beigesetzt deren Sequenzen komplementär zu den Protein-codierenden Abschnitten des Genoms sind. Die Enden dieser Oligos sind mit Biotin gelabelet. Man erlaubt die Hybridisierung der Oligos mit den genomischen DNA Fragmenten und benutzt dann die Biotin Label um die entstandenen Hybirde an streptavidin-dekorierte Beads zu binden. Nun kann man die ungebundenen DNA Fragmente, also Fragmente deren Sequenzen nicht für Protein kodieren, wegwaschen. Denaturiert

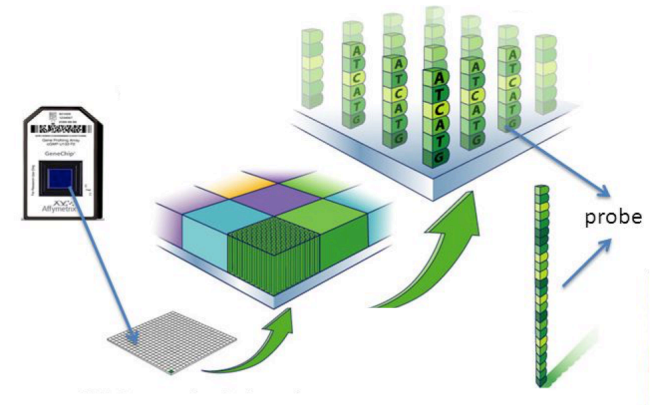

Abbildung 4 *Oberfläche eines Genexperssions-microarrays. Die unterschiedlichen Oligonukleotidcluster sind in unterschiedlichen Farben angezeigt.*

man die DNA nun wiederum, so lösen sich die verbliebenen genomischen DNA Fragmente und können aufgefangen werden. Die so erhaltene Library von protein-kodierenden DNA Fragmenten kann nun mittels Hochdurchsatzsequenzierung (z.B. Illumina *sequencing by synthesis*) untersucht werden.

Genexpressionsstudien mittels DNA Sequenzierung und Microarrays

Neben Mutationen welche die proteinkodierenden DNA Sequenzen direkt verändern, rufen vor allem Mutationen welche die Expressionsmuster von Genen beeinflussen besonders häufig starke phänotypische Veränderungen hervor. Viele genetische Studien untersuchen daher wie Veränderungen in der DNA Sequenz die Genexpressionsmuster eines Organismus verändern.

Diese Art von Genexperssionsstudien wurde in den 1990 Jahren durch die Einführung von Genexperessions-microarrays (Abbildung 4) revolutioniert. Vor der Einführung dieser Microarrays musste der Grad der Genexpression für jedes einzelne Gen in einem separaten, recht zeitaufwendigen, *Northern Blot* Experiment bestimmt werden. Microarrays machten es nun möglich die Expression aller Gene eines Organismus in einem einzigen Experiment zu untersuchen.

Genexpressionsmicroarrays basieren, wie Genotypisierungsmicroarrays auch, auf an einer Oberfläche immobilisierten Clustern von Oligonucleotiden. Die Sequenzen dieser Oligos sind hier aber so gewählt, dass sie den Sequenzen der zu untersuchend mRNAs entsprechen. Um das Genexpressionsniveau in einer Probe zu bestimmen extrahiert man die in der Probe enthaltenen mRNAs und synthetisiert mittels einer reversen Transkriptase die dazu komplementären cDNAs (cDNAs sind chemisch stabiler und daher experimentell einfacher zu handhaben). Diese cDNAs werden anschliessend mit Fluoreszenzmarkern kovalent gelabelt. Gibt man nun diese gelabelten cDNAs auf das Microarray so hybridisiert jede cDNA Spezies bevorzugt mit den zu ihr komplementären Oligos. Auf diese Weise werden die cDNAs an der Position auf dem Array angereichtet, an der sich die zu ihnen komplementären Oligos befinden. Die Stärke der Genexpression eines bestimmten Gens lässt sich nun an der Intensität des Fluoreszenssignals an der Position des entsprechenden Oligo clusters ablesen.

Aus technischen Gründen ist die Bestimmung relativer Genexperssionslevel durch Genexpressionsmicroarrays wesentlich zuverlässiger, als deren absolute Bestimmung. Normalerweise verwendet man daher Genexpressionsmicroarrays zum Vergleich der Genexpression zwischen zwei Proben (Abbildung 5).

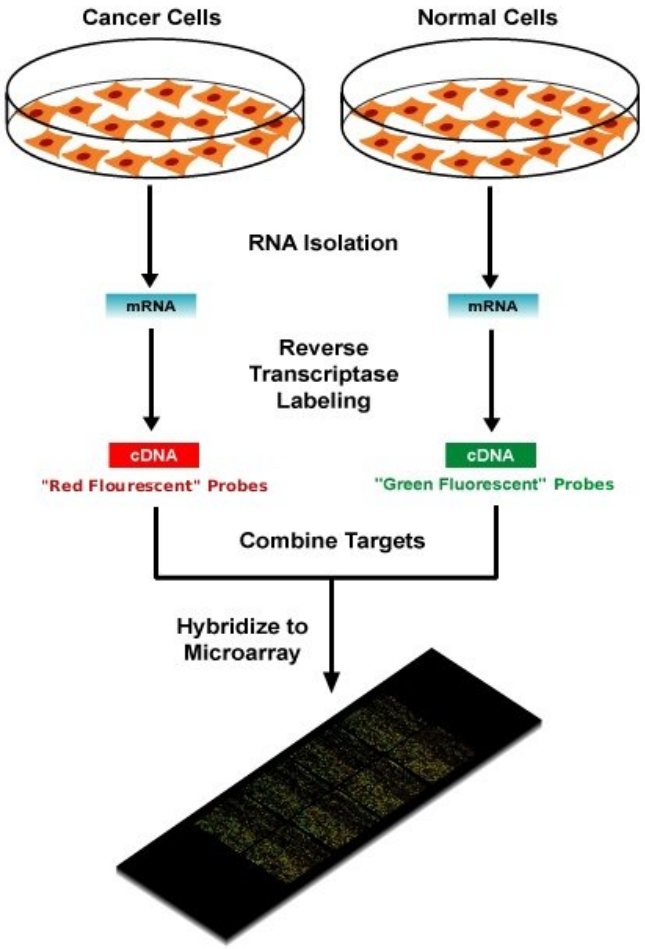

Abbildung 5 *Strategie zur Bestimmung des relativen Expressionsniveaus zwischen zwei Zelllinien. Aus beiden Proben werden die mRNAs extrahiert, in cDNAs reverse-transkribiert und dann mit Fluorophoren gelabelet. Dabei verwendet man für die zwei Proben unterschiedliche Fluorophore (rot und grün). Die Proben werden dann gepoolt und auf das Miroarray aufgebracht. Dort binden Sie spezifisch an die zu ihnen komplementären Oligocluster. Durch Vergleich der Intensität des roten und grünen Fluoreszenzsignals für jede Clusterposition lässt sich so für alle Gene bestimmen, ob die unterschiedlichen Bedingungen in den beiden Proben (hier Krebszellen vs. gesunde Zellen) zu einer Veränderung in der Expression geführt hat.*

Inzwischen ist die Verwendung von Microarrays für Genexpressionsstudien fast schon wieder veraltet und wird zunehmend durch die RNA-Seq Technologie abgelöst. Bei der RNA-Seq Technik verwendet man eine reverse Transkriptase um aus den in einer Probe enthaltenen mRNA Molekülen die ihnen entsprechenden cDNAs zu synthetisieren. Der so erzeugte cDNA Pool wird nun durch Hochdruchsatzsequenzierung analysiert. Die generelle Methodik ist dabei dieselbe, die auch bei der Shotgun Sequenzierung verwendet wird. Nur besteht in diesem Fall die Sequenzierlibrary nicht aus Fragmenten der genomischen DNA, sondern aus einem Pool von cDNA Molekülen. Die bei der Sequenzierung generierten Reads werden in der anschliessenden Datenanalyse auf die Genomsequenz gemapped. Stark exprimierte Gene generieren viele mRNA Moleküle und somit auch viele cDNA Moleküle und Reads. Um das Expressionsniveau verschiedener Gene miteinander zu vergleichen, muss man dann nur die Anzahl der auf die jeweiligen Gene gemappten Reads vergleichen.

Protein-DNA Interaktionen und Chromatinstruktur aus Sequenzierdaten

DNA Sequenzierungsmethoden sind in den letzten Jahren so effizient und kostengünstig geworden, dass sie zunehmend auch zur Beantwortung von Fragestellungen verwendet werden, die nur indirekt mit der Sequenz der

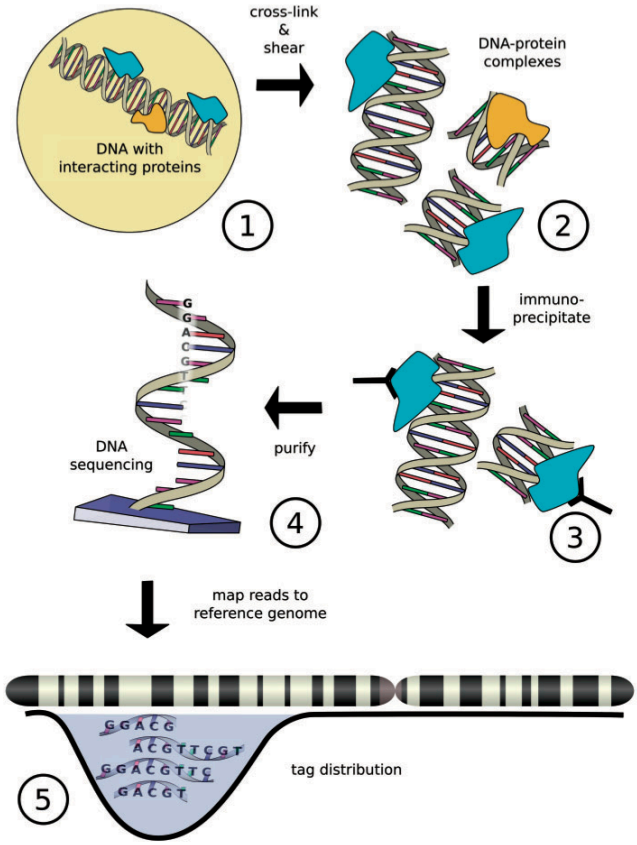

Abbildung 6 *Outline eines „Chromatin-Immunopreciptiation und Sequenzierung“ (Chip-Seq) Experiments. (1) Proteinmoleküle werden durch chemisches Crosslinking mit den an sie gebundenen DNA Abschnitten verknüpft. (2) Die DNA wird anschliessend durch mechanische Behandlung in kurze Bruchstücke fragmentiert. (3) Protein spezifische Antikörper werden nun benutzt, um gezielt solche Protein-DNA Komplexe zu isolieren, die eine bestimmtes Protein enthalten. (4) Diese DNA Fragmente werden sequenziert. (5) Die resultierenden Reads können dann bioinformatisch auf die Sequenz des Genoms gemapped werden und zeigen so wo im Genom dieses bestimmte Protein bindet.*

DNA zu tun haben. Als Beispiel für diese Klasse von Experimenten seien hier *Chromatin Immunoprecipitation and Sequencing* (ChIP-Seq) und *Chromosome Conformation Capture* (3C bzw. Hi-C) genannt.

Ziel eines ChiP-Seq Experiments (Abbildung 6) ist die Bestimmung all der Stellen im Genom, mit denen ein spezifisches Protein (z.B. ein Transkriptionsfaktor, eine Polymerase oder ein Histon) interagiert. Dazu setzt man eine lebendende Zelle dem *crosslinking* Reagenz Formaldehyd aus, welches kovalente aber reversible chemische Verbindungen zwischen benachbarten Proteinen und Nukleinsäuren erzeugt. Die Zellen werden dann aufgebrochen, die DNA wird mechanisch in kleine Stücke geschert. Auf Beads immobilisierte, spezifische Antikörper werden nun benutzt, um das zu untersuchende Protein (zusammen mit den gebundenen DNA Fragmenten) von den anderen Zellbestandteilen zu trennen. Die *crosslinking* Reaktion wird nun rückgängig gemacht und so werden die zuvor an das Protein gebundenen DNA Fragmente freigesetzt. Die Sequenz dieser DNA Fragmente wir nun durch Hochdruchsatzsequenzierung (z.B. Sequenziermethoden der 2. Generation) bestimmt. Durch Sequenzalignmentalgorithmen ist es dann möglich, mittels dieser Sequenzen die Stellen des Genoms zu finden, mit denen dieser Typ von Protein in der lebenden Zelle zum Zeitpunkt des *crosslinkings* assoziiert war.

Eine ähnliche Strategie wird in *chromosome conformation capture* Experimenten wie z.B. einem sogenannten Hi-C Experiment verwendet. Diese Experimente verwendet man aber *crosslinks* zwischen benachbarten DNA-Abschnitten. Nach der *crosslinking* Reaktion wird die DNA mit einem recht unspezifischen Restriktionsenzym in kleine Stücke geschnitten. Die bei diesen Schnitten entstandenen Einzelstrangüberhänge werden mittels einer DNA Polymerase mit biotin-markierten Nukleotiden aufgefüllt. Die so entstandenen *blunt ends* werden dann durch eine Ligase verknüpft. Dabei entstehen lineare DNA Moleküle aus DNA-Abschnitten, die in der linearen Genomsequenz weit voneinander entfernt waren, aber in der gefalteten räumlichen Chromatinstruktur des Zellkerns nah beieinanderlagen. Diese Hybridstränge werden nun noch einmal mechanisch kürzere Stücke zerbrochen und dann über Streptavidin-dekorierte Beads (Streptavidin bindet Biotin sehr stark und spezifisch) aufgereinigt. Diese Hybrid-DNA Fragmente werden sequenziert und die erhaltenen Sequenzen im Computer mit der Genomsequenz der Zelle aligniert. Die zwei Enden dieser Hybridsequenzen alignieren dabei jeweils mit dem Abschnitt des Genoms aus dem sie ursprünglich stammen und zeigen an, dass diese beiden Abschnitte des Genoms im Zellkern physisch benachbart waren.

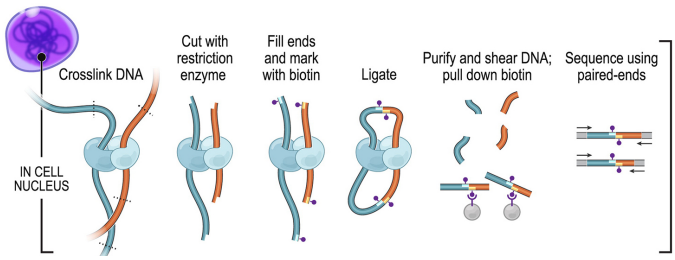

Abbildung 7 *Hi-C Experiment zur Bestimmung der strukturellen Organisation des Genoms in einer Zelle. Durch Crosslinking, Enzymatische Restriktion und Ligation werden ursprünglich separate DNA Stränge (blau und orange,) die in der Zelle physisch nahe bei einander lagen, zu linearen DNA Molekülen verknüpft. Diese Hybridmoleküle werden dann isoliert und sequenziert. Jeder der resultierenden Reads (rechts) enthält zwei Abschnitte (blau und orange) deren genomischer Ursprung jeweils durch Sequenzalignment mit der Genomsequenz bestimmt werden kann. Die Tatsache, dass diese beiden Sequenzabschnitte in einem Read auftreten, zeigt an das die entsprechenden Abschnitte des Genoms im Zellkern physisch nahe bei einander lagen. (Abbildungsquelle: Rao et al. 2014 Cell)*

Dank der hohen Effizienz der DNA Sequenzierungstechnologien ist es möglich auf diese Weise Milliarden von paarweisen Interaktionen zu identifizieren. Basierend auf einem solchen Datensatz wurde z.B. die 3-dimensional Organisation des menschlichen Genoms im Zellkern untersucht und dessen hierarchischen Organisationsprinzipien aufgeklärt.

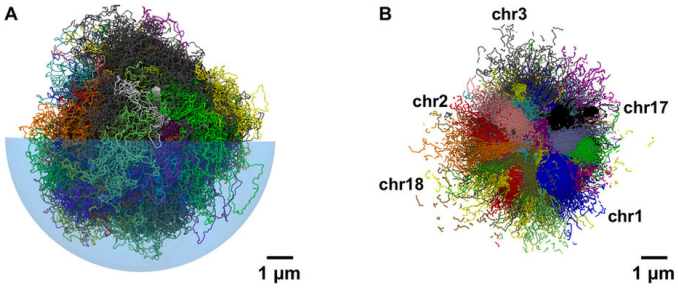

Abbildung 8 *3-Dimensional Rekonstruktion der physischen Organisation des menschlichen Genoms im Zellkern. (A) zeigt eine Aussicht und (B) einen Querschnitt durch diese Rekonstruktion. Die Rohdaten für diese Rekonstruktion lieferten 15 Milliarden mit der Hi-C Technik generierte inter-strang Reads. (Abbildungsquelle: Di Stefano et al.2016 Scientific Reports)*

Von DNA Sequenzierung zu DNA Synthese

Viele der in diesem Handout beschriebenen Techniken (z.B. Microarrays und Exon Sequenzierung), sowie einige der genomweiten Techniken (RNAi & CRISPR) die später im Kurs besprochen werden, basieren auf Libraries von

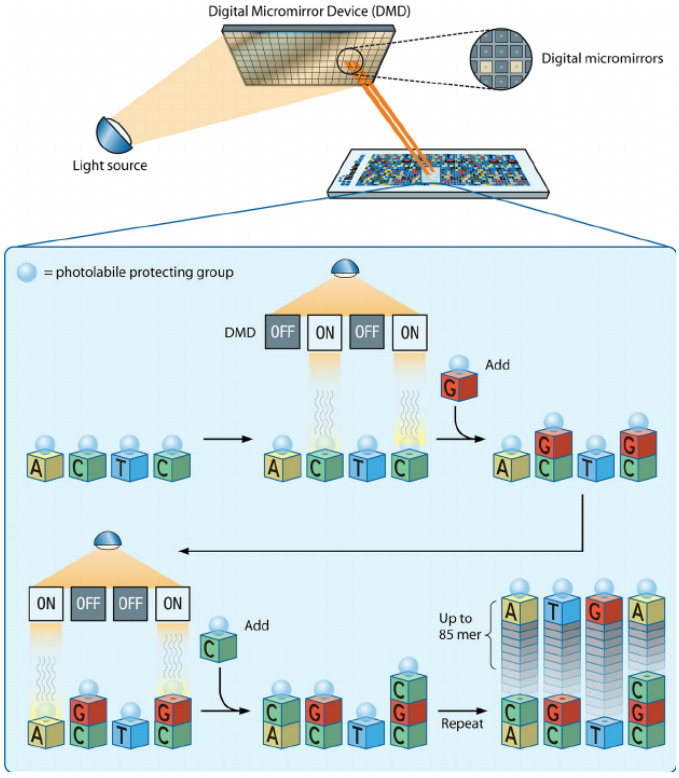

Abbildung 9 *Lichtgesteuerte massiv paralelle Synthese von immobilisierten Oligonucleotiden. Während der Synthese werden gezielt nur diejenigen Bereiche der Oberfläche aktiviert an denen im nächsten Reaktionsschritt ein bestimmtes Nukleotid eingebaut werden soll. Auf diese Weise lassen sich viel tausende Oligos mit freiwählbarer Sequenz auf einem einzelnen Chip synthetisieren. (Abbildungsquelle: Nimblegen)*

Tausenden oder sogar Millionen verschiedener Oligonukleotide (Oligos) mit spezifischen Sequenzen. Wie ist es möglich eine derartige Anzahl von Oligonkleotiden im Labor herzustellen?

Wie bei der Sequenzierung ist auch hier wieder Miniaturisierung und Parallelisierung sowie eine kontinuierliche Optimierung der Reagenzien und Reaktionsbedingungen der Schlüssel.

Dabei sind die Oligos, ähnlich wie bei den DNA Sequenzierungsmethoden der 2. und 3. Generation, während der Synthese auf einer Oberfläche immobilisiert. Die Identität eines jeden zu synthetisierenden Oligos ist also durch dessen xy-Postion auf der Oberfläche gegeben. Während der Synthese müssen den wachsenden Oligo Sequenzen an jeder dieser xy-Positionen nun die Nukleotide entsprechend der vorher festgelegten DNA Sequenz hinzugefügt werden. Verschiedene Firmen haben dafür eine Reihe von unterschiedliche Methoden entwickelt. Die Firma Agilent verwendet z.B. Inkjet Technologie wie sie ursprünglich für Tintenstrahl Drucker entwickelt wurde und „druckt“ so in jedem Reaktionszyklus an jeder xy-Position

mit einem winzigen Reagenztröpfchen das jeweils richtig Nukleotid auf die Oberfläche.

Ein alternativer Ansatz (Abbildung 9), der z.B.von den Firmen Affymetrix, Nimblegen und CustomArray entwickelt wurde, ist die gesamte Oberfläche nacheinander mit jedem der vier Nukleotide „überfluten“. Die Reaktion dieser Nukleotide mit den wachsenden Oligonucleotidmoleküle wird dann aber nur an denjenigen Positionen aktiviert (also zur Reaktion befähigt), an denen dieses neue Nukleotid eingebaut werden soll. Nach der Reaktion werden die nicht aktivierten Nukleotidmoleküle weggewaschen und der Prozess wird mit den drei anderen Nukleotiden wiederholt. So kann man in jedem Reaktionsschritt an jedem wachsenden Oligonukleotidstrang das jeweils gewünschte Nukleotid einfügen. Die positionsspezifische Aktivierung der Nukleotidmoleküle findet dabei entweder durch eine photo- oder elektrochemische Entfernung von speziell für diesen Zweck entwickelten Blockergruppen statt. Für die photochemische Aktivierung verwendet man dabei computergesteuerte *micromirror arrays,* die für digitale Projektoren entwickelt wurden und eine punktgenaue Beleuchtung der Reaktionsoberfläche erlauben. Für die elektrochemische Aktivierung werden als Syntheseoberflächen spezielle Chips mit eingebetteten digital ansteuerbare Elektroden verwendet, welche den pH-Wert auf der Oberfläche lokal absenken können und dadurch die säureempfindliche Blockergruppen der Nukleotidemoleküle entfernen.

Allen drei Ansätzen ist gemein, dass sie die parallelisierte Synthese von im Computer entworfene Oligonukleotidsequenzen ermöglichen. Durch die Automatisierung und Miniaturisierung fallen die Kosten für solche Synthesen überraschend moderat aus. Diverse Firmen bieten die Synthese von 100'000 unterschiedlichen Oligos mit vom Nutzer freiwählbaren Sequenzen von jeweils 150 Basen Länge für knapp über $5000 an (Stand Frühjahr 2017).

Die derzeit limitierenden Faktoren für die Verwendung dieser synthetischen Oligos sind die Länge der Sequenzen die routinemässig synthetisiert werden können (je nach Technologie zwischen 75 und 200 Basen) und die nicht unerhebliche Rate (~0.1%) mit der einzelne Basen falsch (bzw. nicht) eingebaut werden.

Automatisierte Methoden mit der solche synthetischen Oligos zu längeren Sequenzen verknüpft und etwaige Fehler entfernt werden können sind zurzeit in der Entwicklung und erfahren eine rapide Verbesserung. Es scheint daher absehbar, dass die komplette Synthese von Genen und Genomen in den kommenden Jahren ähnlich selbstverständlich wird, wie es deren Sequenzierung bereits heute ist.