Einführung in NGS & Exomsequenzierung

Peter N. Robinson

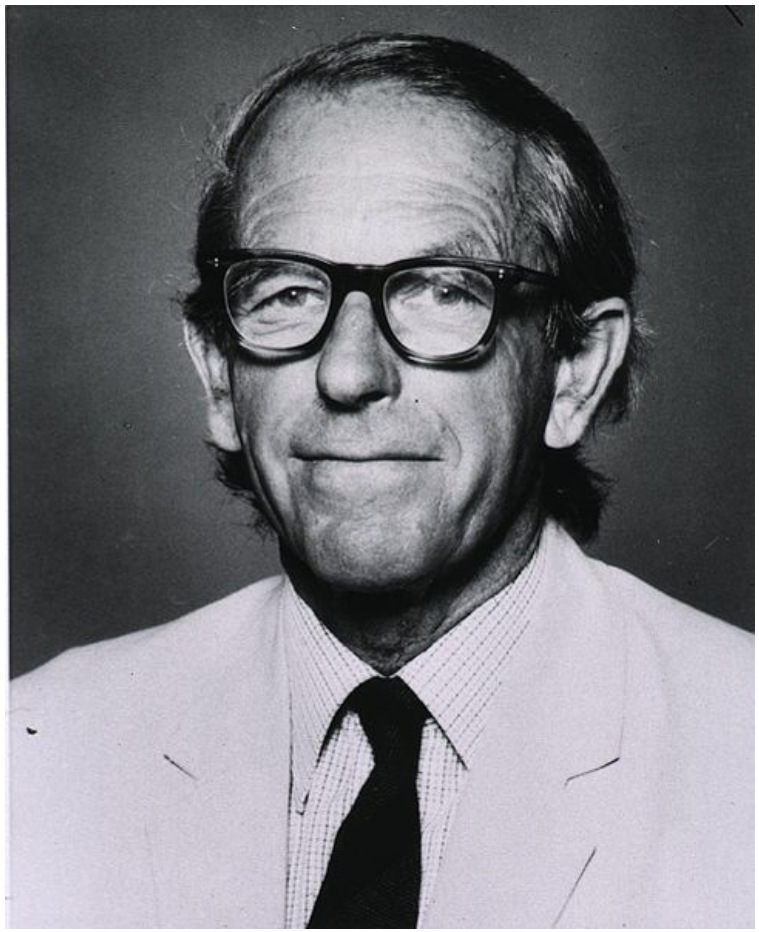
Institut für medizinische Genetik
Charité Universitätsmedizin Berlin

26. Januar 2015

Outline

Fred Sanger: $1\frac{1}{4}$ Nobelpreise

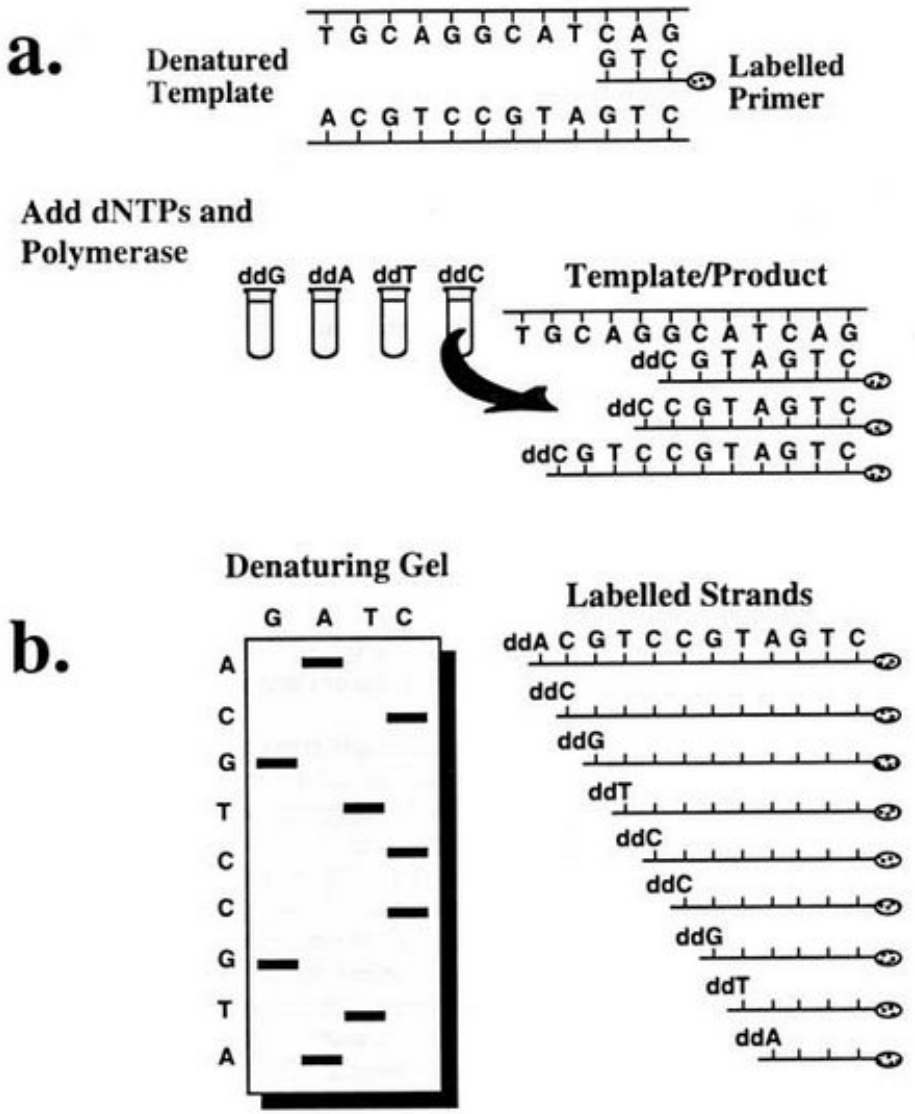
- 1958: Nobelpreis für Chemie "für die Aufklärung der Insulin-Struktur und seine Arbeiten zur Protein-Sequenzierung".
- 1980, Nobelpreis für Chemie ($\frac{1}{4}$) "für Untersuchungen zur Ermittlung der Basensequenz in Nukleinsäuren.".
- Sangersequenzierung: Bis vor kurzem die Standardmethode zur Ermittlung von DNA-Sequenzen



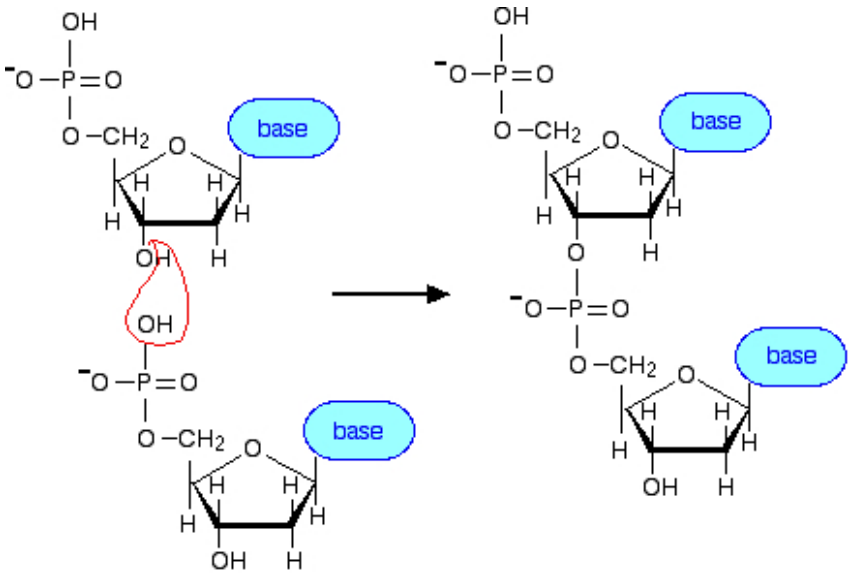
1918–
British biochemist

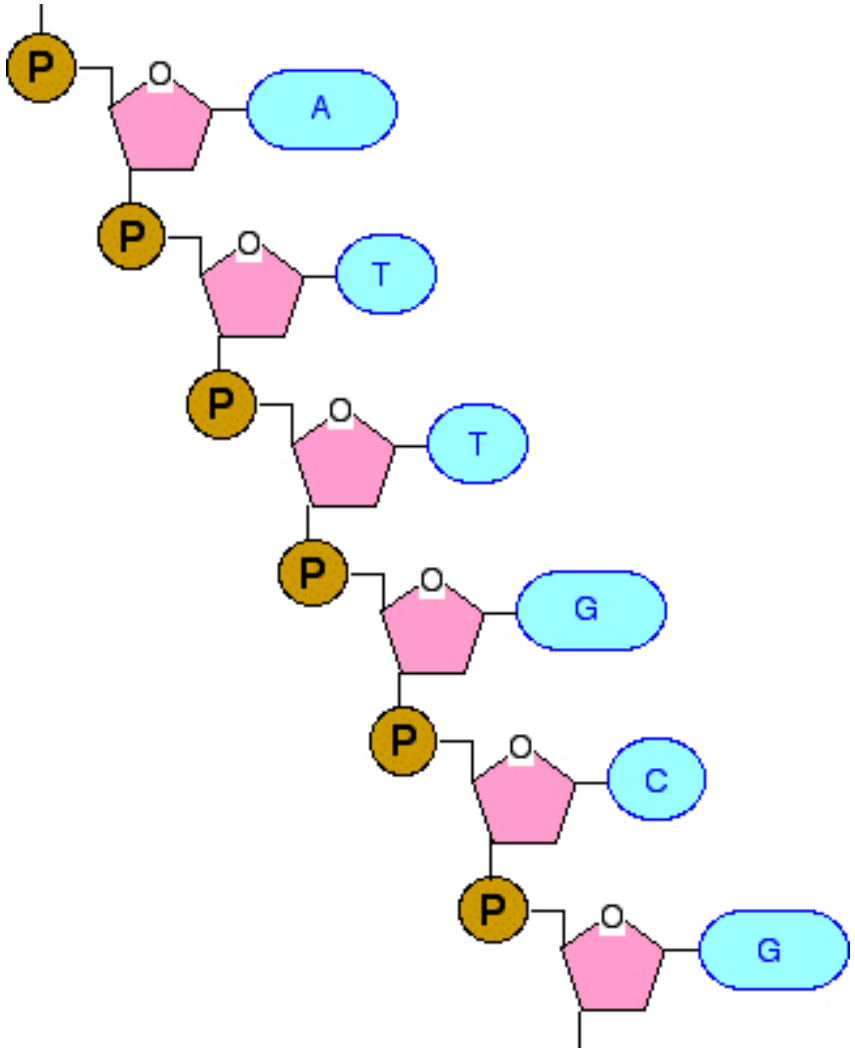
Sangersequenzierung

- Kettenabbruchmethode (Didesoxy-ddNTPs)
- “Zutaten”:
  1. DNA-Matrize.
  2. DNA-Primer
  3. DNA-Polymerase
  4. normale Desoxynukleosidtriphosphat A,C,G,T (dNTP)
  5. Kettenabbruch-ddNTPs



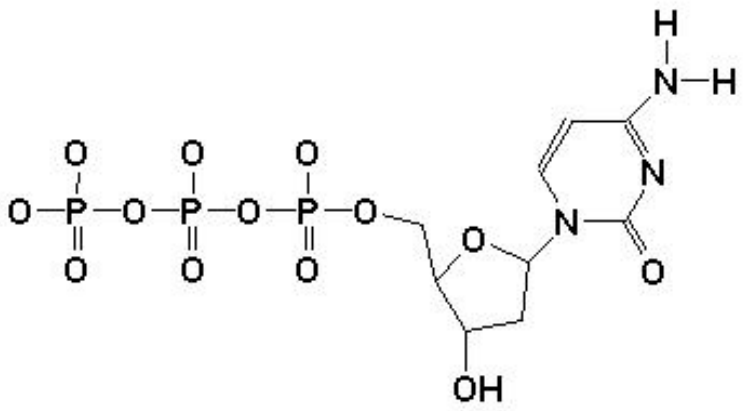
DNA Synthese: Chain extension



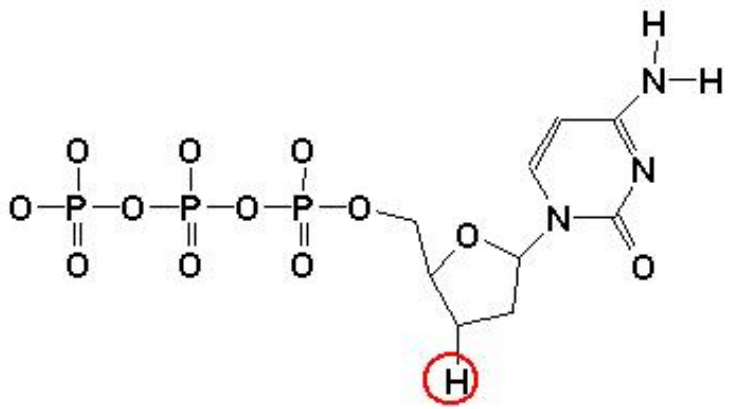


- DNA wird von 5’ nach 3’ verlängert

Sangersequenzierung: Kettenabbruch-ddNTPs



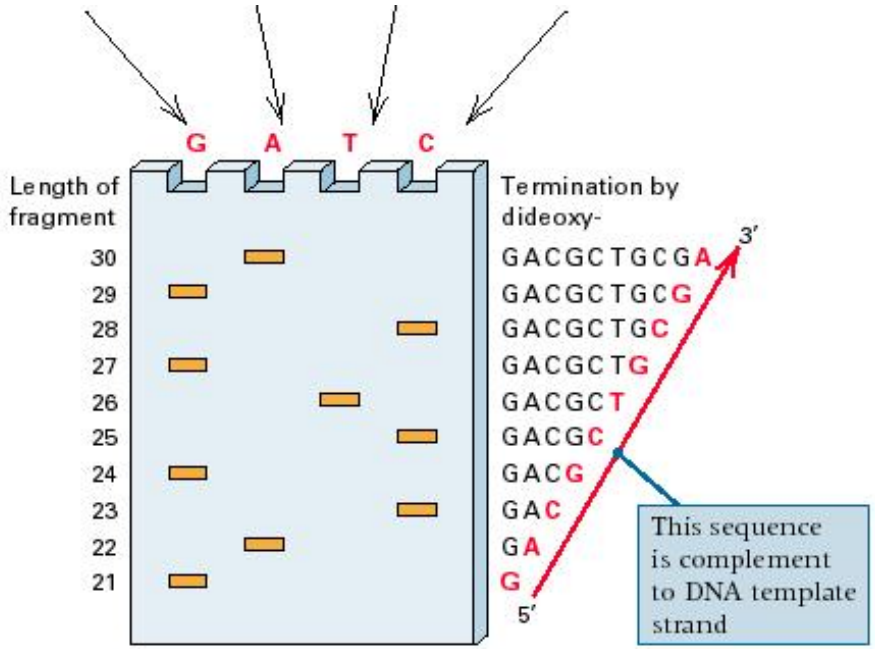
Desoxycytosin (dCTP)

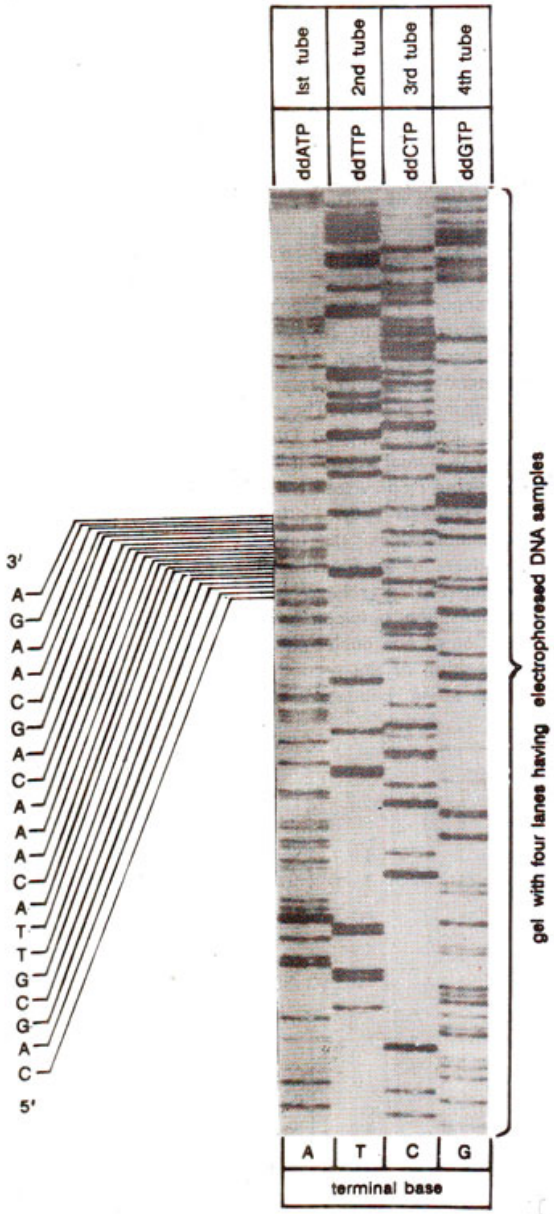


Didesoxycytosin (ddCTP)

- Diese Kettenabbruch-ddNTPs besitzen keine 3'-Hydroxygruppe: Werden sie in den neusynthetisierten Strang eingebaut, ist eine Verlängerung der DNA durch die DNA-Polymerase nicht mehr möglich, da die OH-Gruppe am 3'-C-Atom für die Verknüpfung mit der Phosphatgruppe des nächsten Nukleotids fehlt.

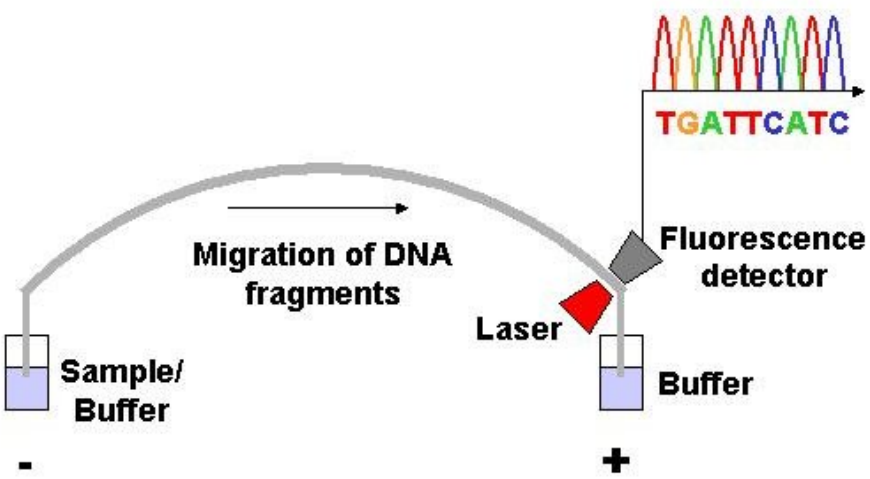
Sangersequenzierung: Radioaktiv

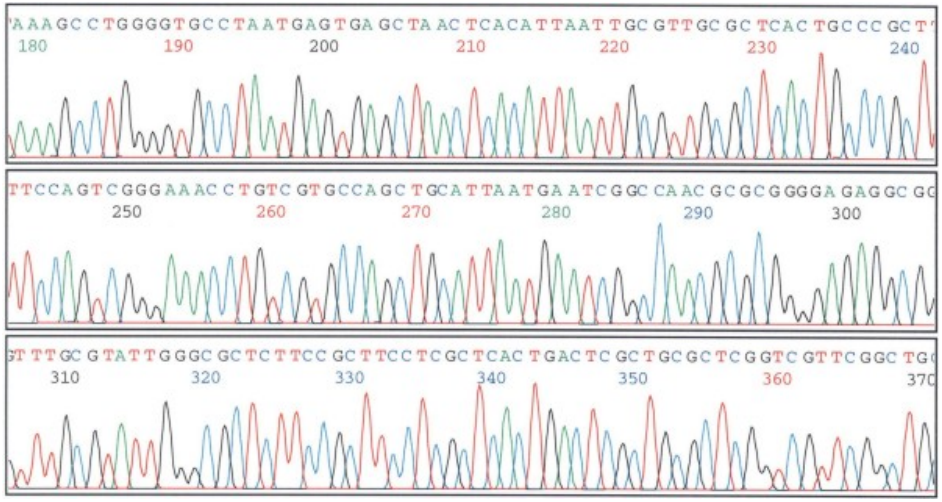




- radioaktiv markierte Nukleotide, z.B., dATP–[α-^{33}P]
- oder markierte Primer, vier Reaktionen (eine für jedes ddNTP)

Sangersequenzierung: Fluoreszent





- “Dye-terminator” Sequenzierung
- jedes ddNTP wird mit einem unterschiedlichen fluoreszenten Farbstoff markiert (unterschiedliche Wellenlänge)
- Daher nur eine Reaktion notwendig
- Intensität jeder Wellenlänge wird gegen die elektrophoretische Zeit geplottet (“chromatogram”)
- Farben: A, T, C, G

Sangersequenzierung: HGP

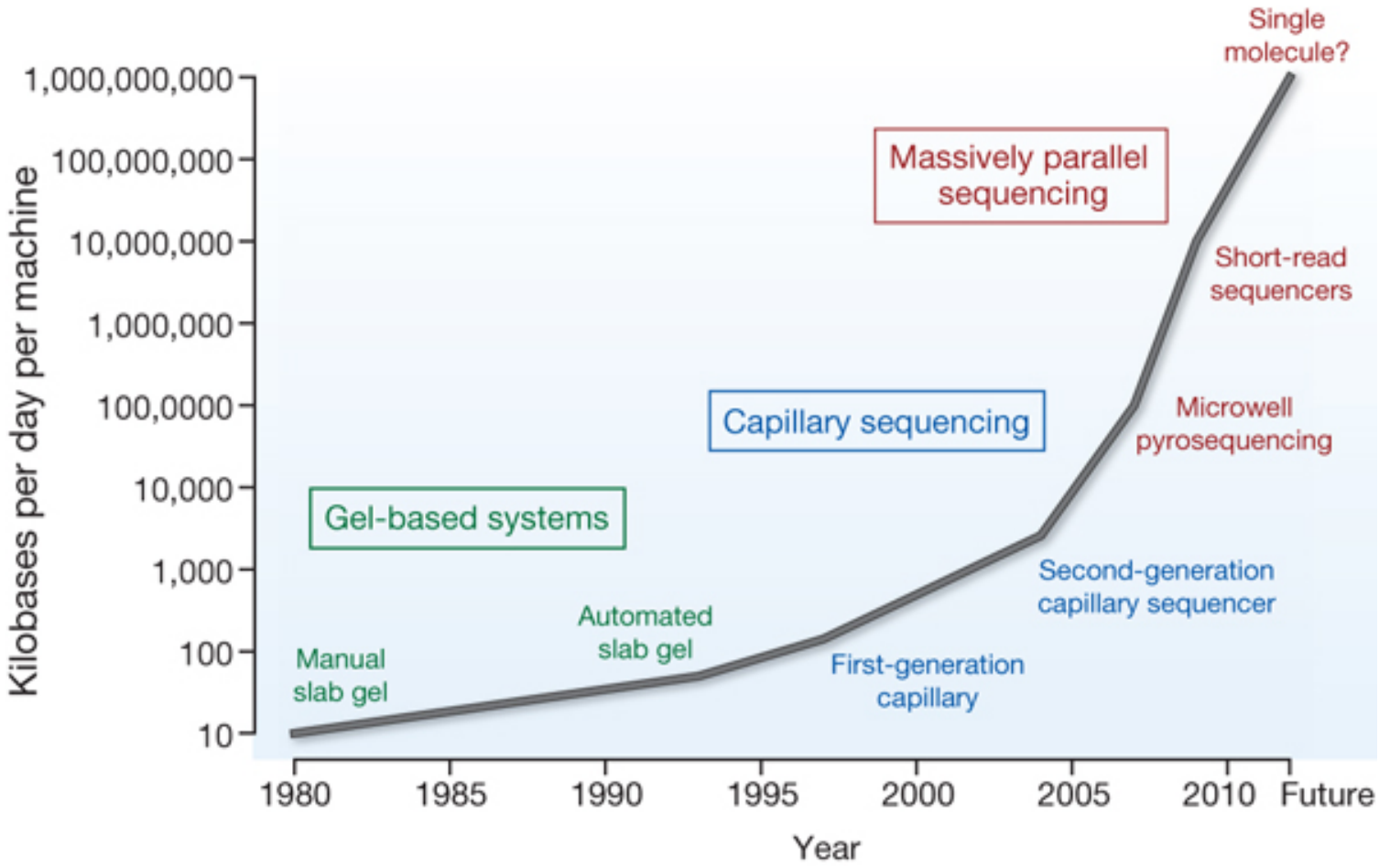
- Sangersequenzierung ermöglichte die erste Charakterisierung des humanen Genoms
- Aber: Beschränkter Durchsatz
- In der Glanzzeit der Sangersequenzierung, 400 kb pro Machine pro Tag
- ca. 45.000 Läufe für ein humanes Genom (6x) ...

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome

Nature **409**:860-921

Outline

Next-Generation Sequencing



MR Stratton et al. *Nature* **458**, 719-724 (2009)

- NGS: verschiedene Technologie, welche eine massive Parallelisierung der DNA-Sequenzierung ermöglichen

Next-Generation Sequencing (NGS)

- Genbank 2005 – 50 Gb Daten
- Illumina GA: 1000 Genomes-Project im Jahr 2008 – 2,500 Gb
- “Each week in Sept–Oct of 2008, the 1000 Genomes Project created the equivalent of all the data in GenBank”

Thomas Keane and Jan Aerts. Tutorial 1: Working with next-generation sequencing data - A short primer on QC, alignment, and variation analysis of next-generation sequencing data. 9th European Conference on Computational Biology 26th September, 2010

Illumina Sequencing

- Mehrere konkurriende NGS-Platformen
- Diejenige von Illumina scheint momentan für die meisten Applikationen überlegen zu sein
- Vier grundlegende Schritte:

1. DNA & Library Präparation	Fragmentierung der DNA und Anfügung von Adaptoren
2. Chip/flowcell Präp	DNA-Fragmente an Flowcell anheften, amplifizieren (colony PCR)
3. Sequenzierung	Massiv parallele DNA Sequenzierung
4. Bioinformatische Analyse	verschieden

Library Präp: Adapterligation

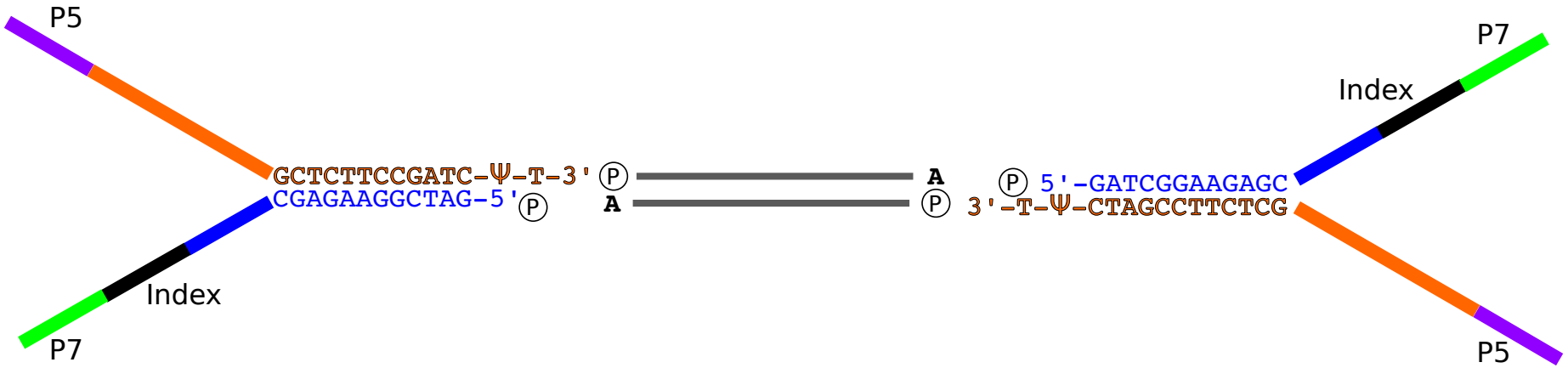
- Erster Schritt: Fragmentierung der DNA-Probe gefolgt von Adaptorligation[1]

Zielstellung der Adapterligation: Spezielle Adaptoren werden an die DNA-Fragmente der Library angefügt (ligiert), was drei Zwecken dient:

1. Molekulare Indizierung (Barcoding) von Proben
2. Spezifische PCR-Anreicherung der DNA-Fragmente der Library
3. Im nachfolgenden Schritt die Bindung der Adaptoren an die Flowcell

[1] Einige biochemische Schritte werden hier übersprungen

Library Prep (3): Adapter ligation



- DNA-Ligation: DNA-Ligase ist ein Enzym, das durch die Bildung einer Phosphodiästerbindung zwei DNA-Fragmente miteinander verbindet
- Wir verwenden DNA-Ligase, um die NGS-Adaptoren an die Fragmente der DNA-Library zu verbinden

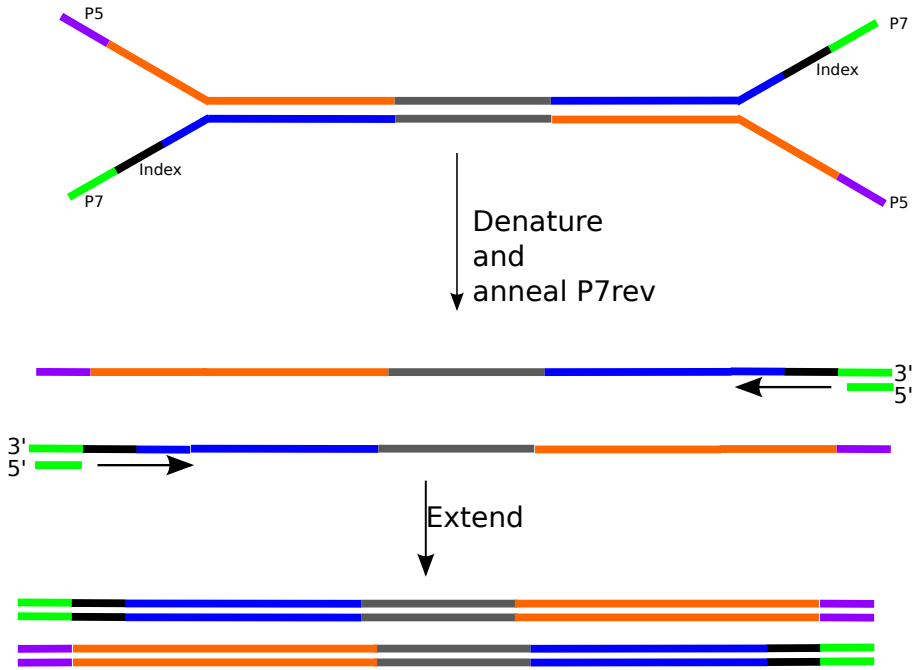
Library Präp: Anreicherungs-PCR

Zielstellung der Anreicherungs-PCR (*enrichment PCR*):

- Spezifische PCR-Anreicherung der DNA-Fragmente der Library
- die Menge an DNA in der Library vermehren

- PCR wird mit Primern durchgeführt, welche sich an die Sequenzen der Adaptoren anlegen (*annealing*)
- Geringe Anzahl von PCR-Zyklen (10), damit die Verteilung der in der Library vertretenen Sequenzen nicht verzerrt wird

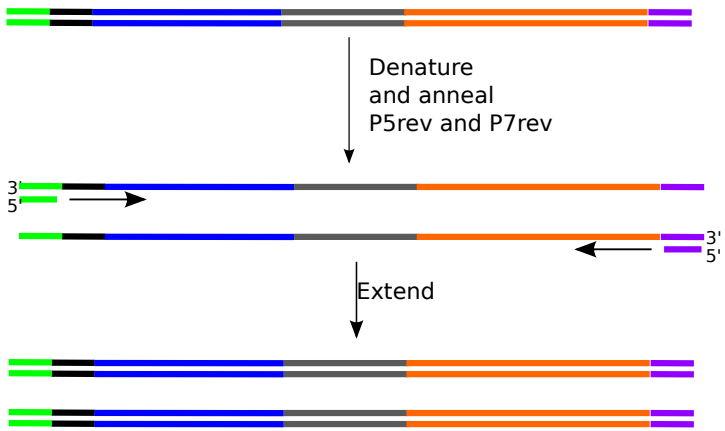
Library Präp: Anreicherungs-PCR



- Der P7-Primer enthält eine Sequenz, die zu den letzten 24 Nukleotiden des Adaptors revers komplementär ist
- 5'-CAAGCAGAAGACGGCATACGAGAT-3'

```
5′-(...)-NNNNNN-ATCTCGTATGCCGTCTTCTGCTTG-3′
.............3′-TAGAGCATACGGCAGAAGACGAAC-5′
```

Library Präp: Anreicherungs-PCR



- In den übrigen Zyklen kann auch der P5 binden (*annealing*)
- P5 ist mit den ersten 44 Nukleotiden des Universaladaptors identisch, und kann somit an dessen durch die PCR erzeugte revers komplementäre Sequenz binden
- 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA-3'

```
5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′
5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA-3′
```

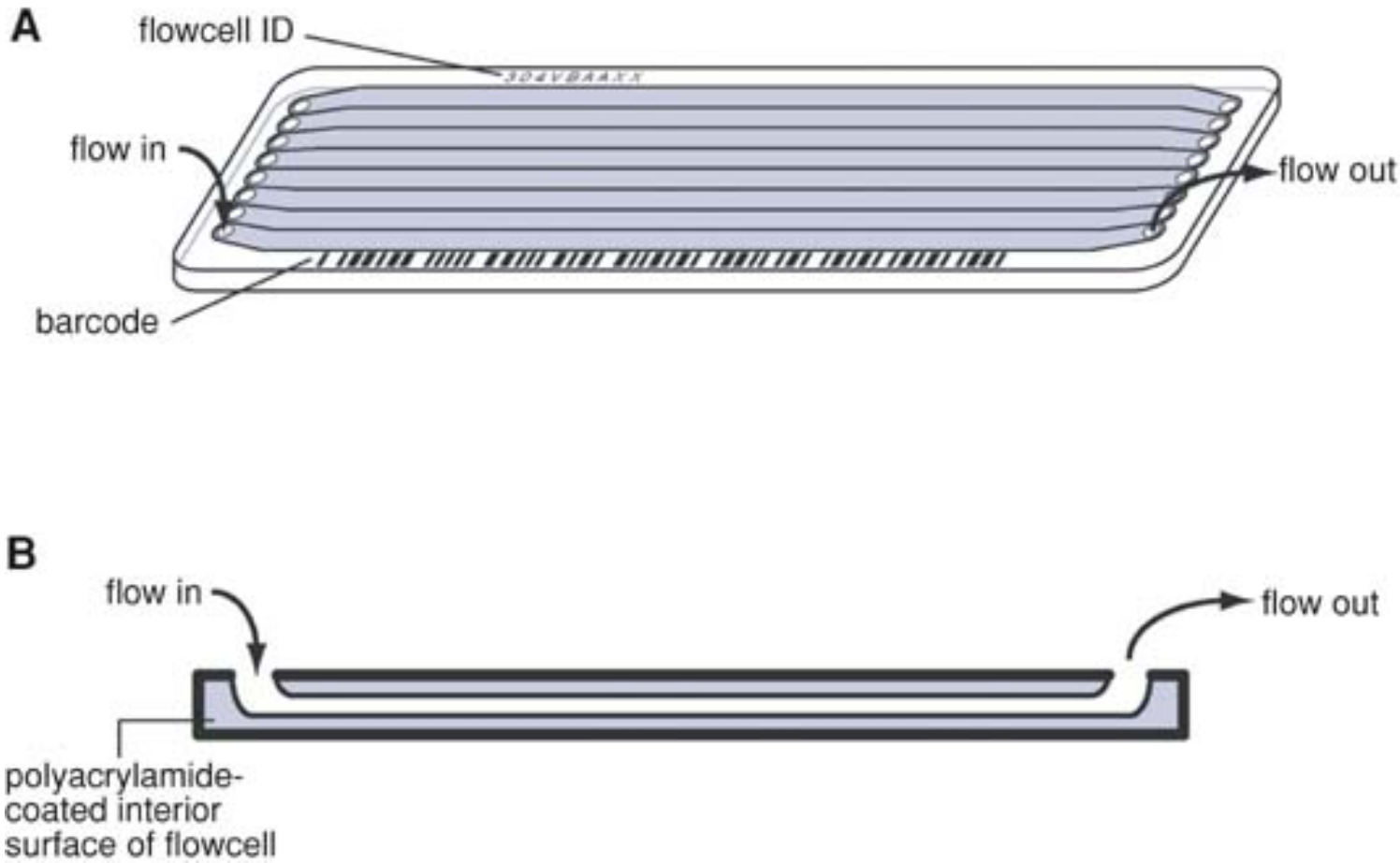
Flow-Cell Präp

Zielstellung der Flowcell Präp:

- Ligierte DNA-Fragmente an die Flowcell binden
-

Flow-Cell Präp

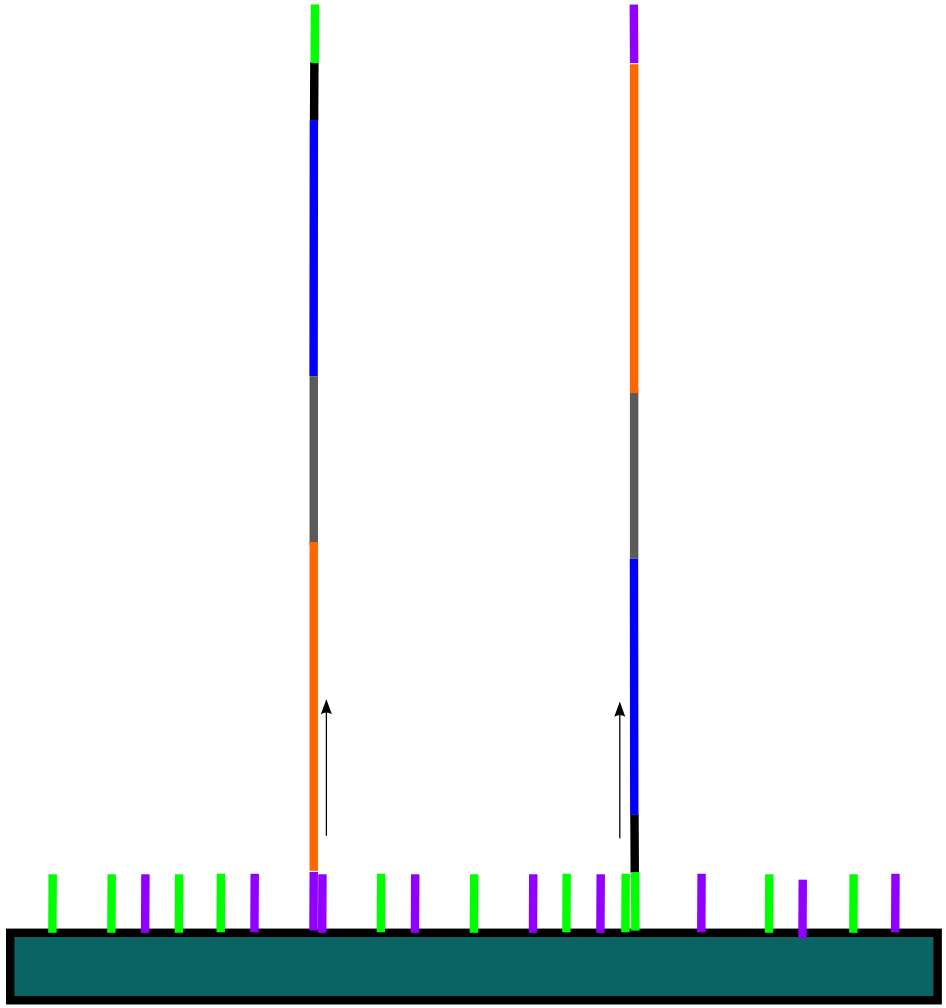
- Eine Flowcell (“Flusszelle”) ist im Prinzip ein Objektträger aus beschichtetem Glaß mit 8 Kanälen



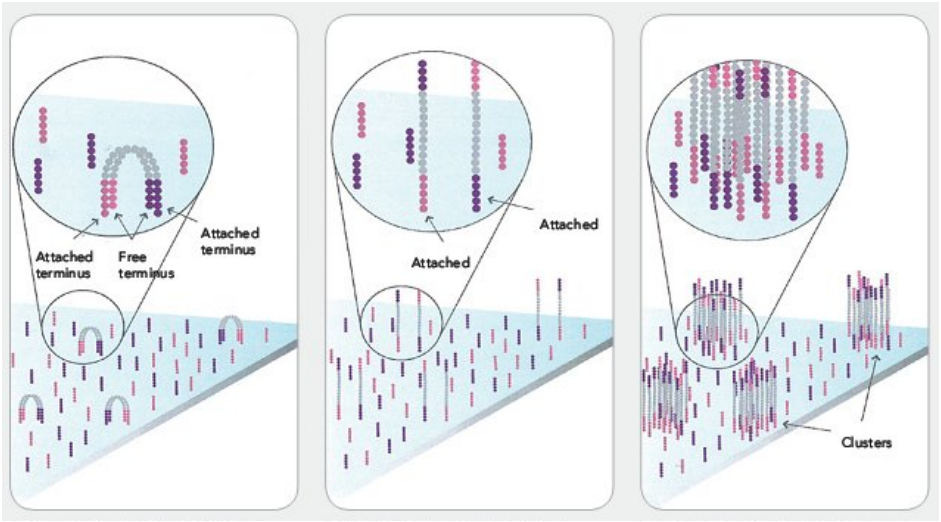
Flow-Cell Präp – Library-Ablagerung

Library deposition

- Extensionsgemisch (Puffer, dNTP's, Taq-Polymerase) wird in die Kanäle der Flowcell gepumpt
- Die Oligos an der Oberfläche der Flowcell werden entsprechend der ligierten DNA-Fragmente verlängert



Flow-Cell Präp – Brückenamplifikation



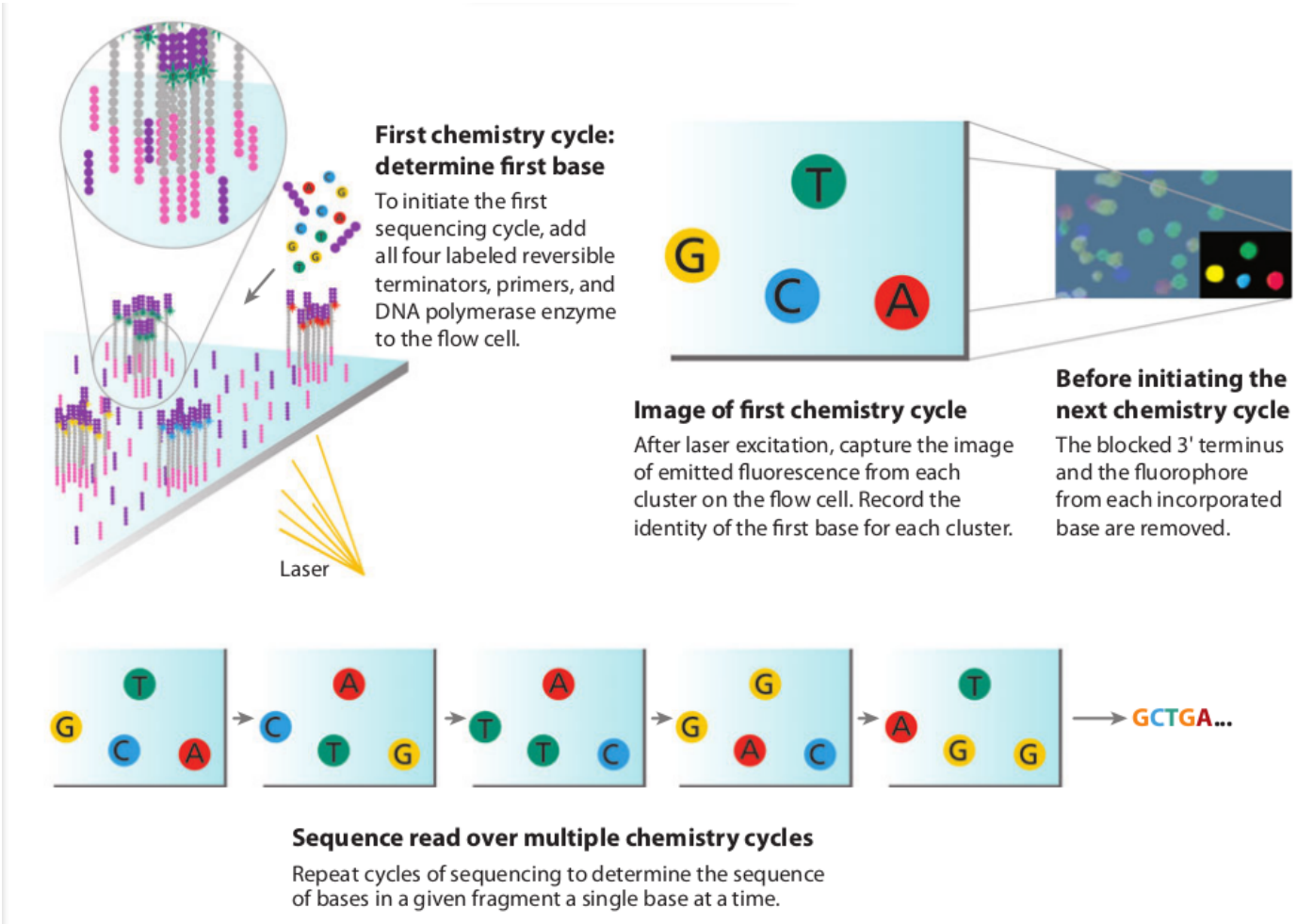
- PCR-Amplifikation an der Oberfläche der Flowcell, “bridge amplification”: (60°C) für 35 Zyklen:
  1. Formamide at 60°C ≈ “Denaturation”
  2. Extensionspuffer ≈ “annealing step”
  3. Extensionsgemisch ≈ “Extension (Verlängerung)” der “normalen” PCR

Sequenzierung durch Synthese

- Sequenzierung durch Synthese (*Sequencing by synthesis*; SBS)
  1. Pro Zyklus wird nur eine Base angefügt (4 markierte ddNTPs)
  2. Unterschied zu Sangersequenzierung: Die ddNTPs haben *reversible* Terminatoren
  3. Nach jedem Zyklus wird die jeweils angefügte Base durch die Bestimmung der spezifischen Wellenlänge der eingebauten ddNTP gemessen
  4. Aufhebung der Blockierung
  5. $\Rightarrow$ Zyklus i+1

Sequenzierung durch Synthese

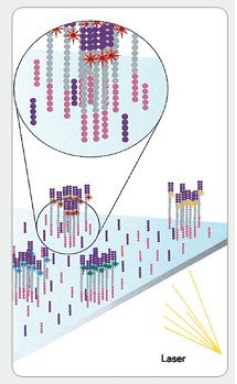
- Sequencing by synthesis:



Mardis E (2008) *Annu Rev Genomics Hum Genet* **9**:387–402

Illumina: Base-Calling

- Base-calling Algorithmen weisen jeder Position ein Nukleotid und einen Qualitätswert zu



Die Qualität wird durch verschiedene Parameter beeinflusst:

- PCR Fehler bei der Kolonie-Amplifikation
- Phasenfehler (Bestimmte Stränge bauen in einem bestimmten Zyklus kein Nukleotid ein und hängen hinter anderen Strängen nach)
- Unreinheiten auf der Flow cell

Die Qualität der Basenzuweisungen (base calls) wird mit dem **PHRED**-Score angegeben

FASTQ und PHRED-Qualitätsscores

- FASTQ-Format.

```
@My-Illu:6:73:941:1973#0/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

1. Read-ID
2. die Sequenz
3. '+' (optional Beschreibung der Sequenz)
4. ASCII-kodierte PHRED-Scores für die entsprechenden Basen

PHRED-Scores

- Der PHRED-Score ist definiert als

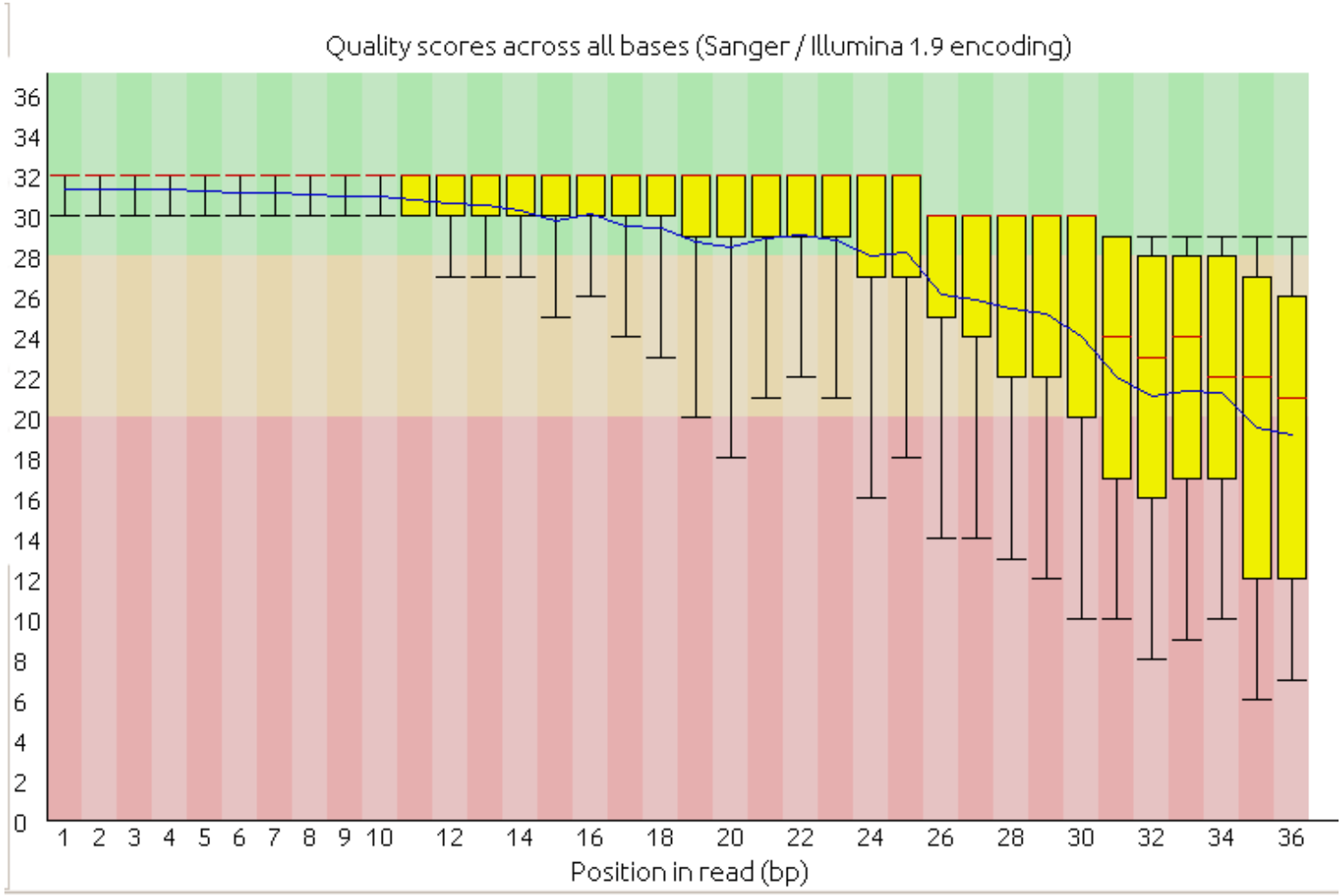
$$\boxed{Q_{PHRED} = -10 \log_{10} p} \tag{1}$$

wobei p die Wahrscheinlichkeit angibt, dass die entsprechende Basenzuweisung ("base call") falsch ist.

Q_{PHRED}	p	Fehlerfreiheit
10	10^{-1}	90%
20	10^{-2}	99%
30	10^{-3}	99.9%
40	10^{-4}	99.99%
50	10^{-5}	99.999%

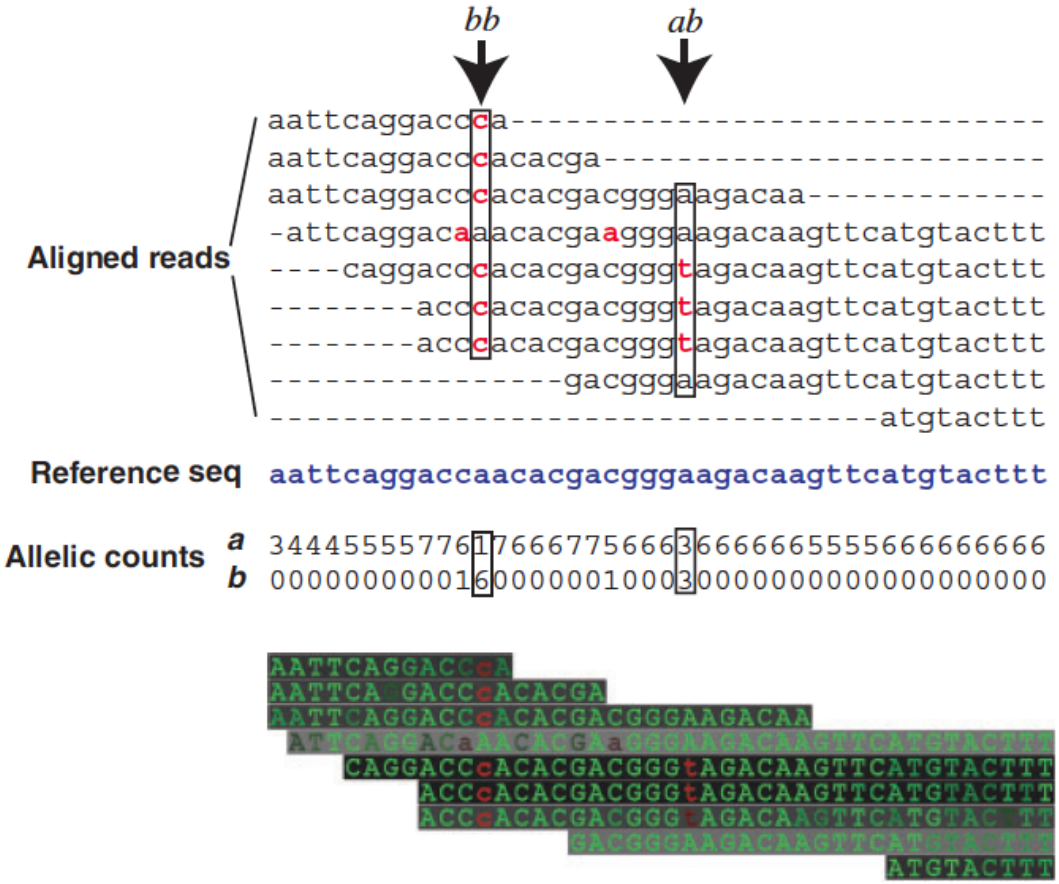
PHRED-Score: Beispiel

- Medianwerte (rot) und Durchschnittswerte (blau) für PHRED-Qualitätsscores bei Illumina 1G (alt!) Daten



Variant-Calling

- Base calls & Phred scores
- Mapping quality
- Alignment



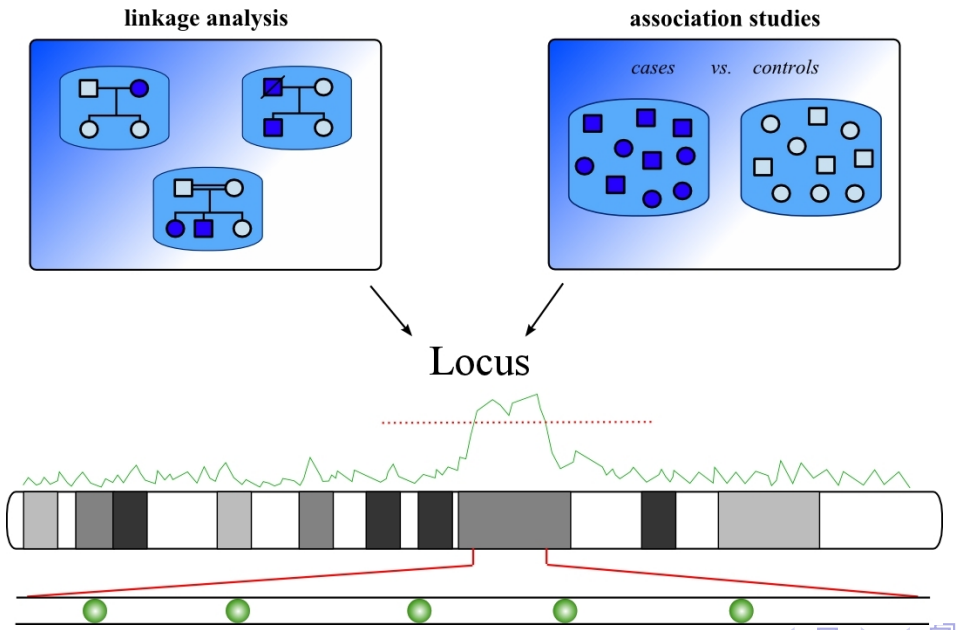
Dunkler Hintergrund ⇔ gute Mapping-Qualität Hoher Kontrast ⇔ hohe Basenqualität

Seltene Erkrankungen

- Häufigkeit in der Bevölkerung $< 1 : 2000$ Personen
- ca. 6% der Bevölkerung hat jeweils eine bestimmte seltene Erkrankung
- Wichtige Subklasse der seltenen Erkrankungen: Mendel'sche (monogene) Erkrankungen

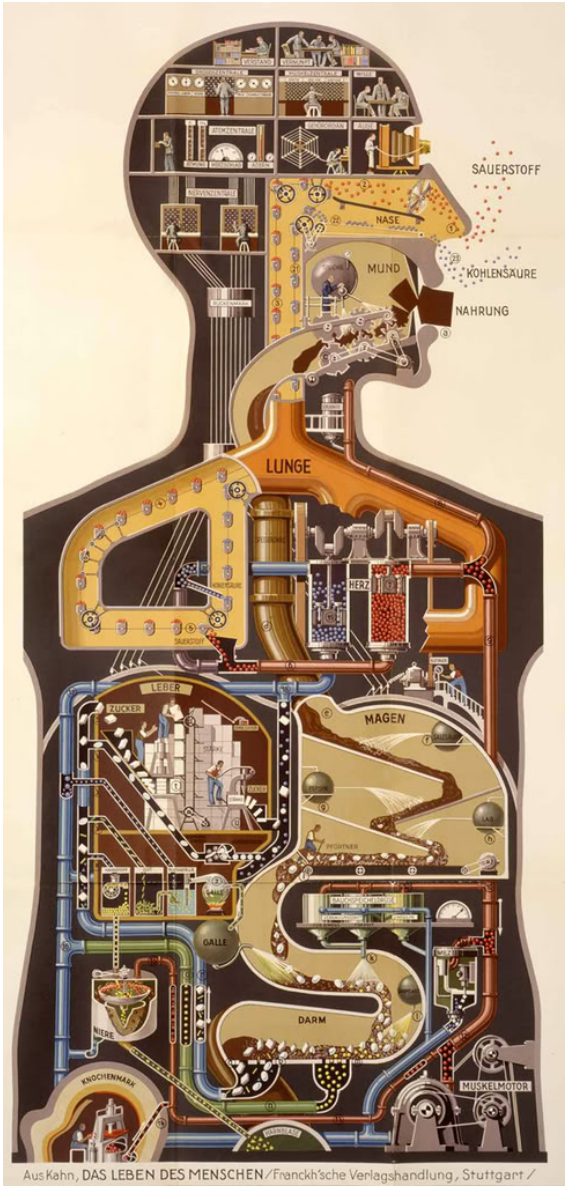
Mendelian disease	
Gen bekannt	2835
Gen unbekannt	1777
Vermutete SE	1989

$\therefore \quad \geq 3766$ Krankheitsgene bleiben zu entdecken

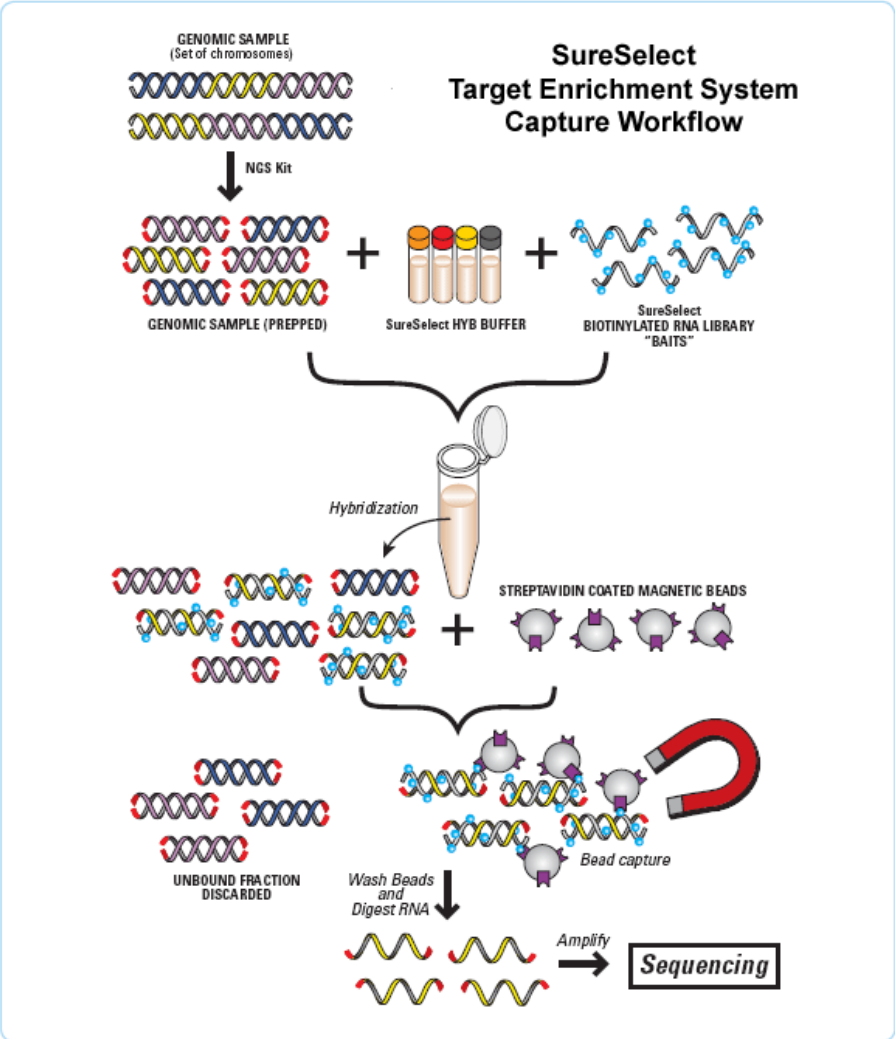


Das Exom

- 249 730 Exons von 24 714 Genen
- Die meisten Mutationen bei Mendel'schen Erkrankungen betreffen das Exom
  1. Nonsense-Mutationen
  2. Missense-Mutationen
  3. Spleiß-Mutationen
  4. Insertionen/Deletionen



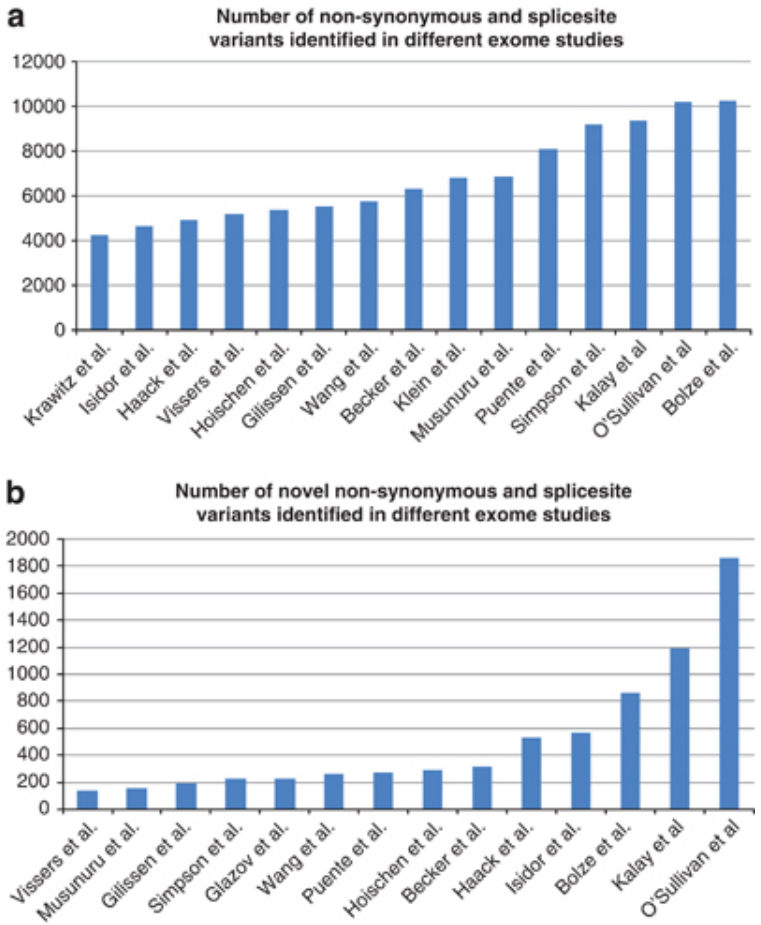
“Capture”-(Anreicherungs)-Verfahren



- Agilent’s SureSelect Exome Enrichment System

die Nadel finden...

- Typisches Ergebnis einer Exomsequenzierung: 40000 oder mehr Varianten
- Häufige und eher nicht pathogene Varianten können herausgefiltert werden, aber es bleiben typischerweise Hunderte bis zu über Tausend Varianten



Gilissen C et al (2012) Disease gene identification strategies for exome sequencing *European Journal of Human Genetics* **20**:490–497

Filtern von Exomdaten

- Die Exomsequenzierung identifiziert typischerweise $\sim$ 30.000 Varianten in jedem Individuum
- Ca. 10.000 dieser Varianten sind in oder direkt neben Exons gelegen
- Ca. 5.000 dieser Varianten sind Missense, Nonsense, Frameshift, usw.
- Wie können wir bei Exomsequenzierungsprojekten die verantwortlichen Mutationen finden?

Filtern von Exomdaten

Wir führen eine Studie mit n Patienten durch und charakterisieren Varianten in M Genen ($n \approx 10$ und $M \approx 20.000$).

- Wir bilden eine $n \times M$ Matrize, $\mathbf{C}$, wobei das Element C_{ij} die Anzahl von Varianten in Gene j bei Patient i angibt
- Sei X_{ij} eine Kodierung des Genotyps von Gene j bei Patient i
- Für eine autosomal rezessive Krankheit gilt:

$$X_{ij} = \mathbf{I}(C_{ij} \geq 2)$$

(d.h., ein Gen muss mindestens zwei Varianten aufweisen, um als Kandidat für eine autosomal rezessive Erkrankung infrage zu kommen)

- Für eine autosomal dominante Krankheit gilt:

$$X_{ij} = \mathbf{I}(C_{ij} \geq 1)$$

Filtern von Exomdaten

- Die Exomsequenzierung identifiziert bei einem einzelnen Patienten m Kandidatenmutationen in den M Genen.[2]
- Die Wahrscheinlichkeit, dass eine bestimmte Mutation in einem der M Genen lokalisiert ist, kann eingeschätzt werden als

$$p = \frac{1}{M} \tag{2}$$

- daher kann die per Zufall zu erwartende Anzahl von Mutationen in einem bestimmten Gen j bei einer Gesamtzahl von m Mutationen nach der Binomialverteilung angegeben werden als

$$C_{ij} \sim \mathrm{B}(m, \frac{1}{M}) \quad \text{i.e.} \quad P(C_{ij} = k) = \binom{m}{k} \frac{1}{M}^{k} \left(1 - \frac{1}{M}\right)^{m-k} \tag{3}$$

[2] m ist typischerweise eine Zahl wie 200–500 Varianten.

Filtern von Exomdaten

- Wir interessieren uns für die Statistik

$$T = \sum_{i=1}^{n} X_{ij} \tag{4}$$

- d.h., wir sequenzieren n Patienten. Was ist die Wahrscheinlichkeit, dass T Patienten Kandidatenmutationen in einem Gen allein per Zufall aufweisen[3]?
- Wir konzentrieren uns im Folgenden auf autosomal dominante Gene.

[3] Zum Beispiel, wenn 100 von 100 Patienten mit Krankheit X eine Mutation in Gen Y haben, dass erscheint es sicher dass Y das echte Krankheitsgen ist. Was ist aber wenn 13 von 100 Patienten eine Mutation haben? Ist das mehr als erwartet?

Filtern von Exomdaten

$$
\begin{aligned}
P(X_{ij} = 1) &= P(C_{ij} \geq 1) \\
&= 1 - P(C_{ij} = 0) \\
&= 1 - \binom{m}{0} \frac{1}{M}^0 \left(1 - \frac{1}{M}\right)^{m-0} \\
&= 1 - \left(1 - \frac{1}{M}\right)^m
\end{aligned}
$$

- Definieren[4] wir $q = \left(1 - \frac{1}{M}\right)$

$$P(X_{ij} = 1) = 1 - q^m$$

[4] z.B beträgt $q = 0.99995$ bei $M = 20000$. Bei $m = 200$ haben wir dann $P(X_{ij}) = 0.00995$.

Filtern von Exomdaten

- Faktorisieren wir nun $1 - q^m$

$$1 - q^m = (1 - q)(1 + q + q^2 + q^3 + \ldots + q^{m-1}) \tag{5}$$

- Daher ergibt sich

$$\begin{aligned} P(X_{ij} = 1) &= 1 - q^m \\ &= (1 - q)(1 + q + q^2 + q^3 + \ldots + q^{m-1}) \\ &= \left(1 - \left(1 - \frac{1}{M}\right)\right)(1 + q + q^2 + q^3 + \ldots + q^{m-1}) \\ &= \frac{1}{M}(1 + q + q^2 + q^3 + \ldots + q^{m-1}) \end{aligned}$$

- In den Klammern befinden sich m Ausdrücke mit einem Wert zwischen q^{m-1} und 1.

$$q^{m-1}\frac{m}{M} \leq P(X_{ij} = 1) \leq \frac{m}{M} \tag{6}$$

Filtern von Exomdaten

Für typische Werte ist diese Approximierung sehr gut, z.B. $q^m = 0.985$ mit $m = 300$, $M = 20.000$. Daher haben wir $P(X_{ij} = 1) \approx \frac{m}{M}$ für die Wahrscheinlichkeit unter der Nullhypothese, dass eine Mutation in Gen j auftritt

- Wir haben daher gezeigt, dass die Wahrscheinlichkeit, dass eine Mutation in Gen j auftritt, nach Bernoulli($p \approx \frac{m}{M}$) verteilt ist
- Mit n Patienten haben wir n Bernoullis, d.h. die Binomialverteilung:

$$T \sim \mathrm{B}(n, \frac{m}{M}) \quad \text{i.e.} \quad P(T = k) = \binom{n}{k} \left(\frac{m}{M}\right)^k \left(1 - \frac{m}{M}\right)^{n-k} \tag{7}$$

Ng et al.: Kabuki Syndrome

Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome

Sarah B Ng[1,7], Abigail W Bigham[2,7], Kati J Buckingham[2], Mark C Hannibal[2,3], Margaret J McMillin[2], Heidi I Gildersleeve[2], Anita E Beck[2,3], Holly K Tabor[2,3], Gregory M Cooper[1], Heather C Mefford[2], Choli Lee[1], Emily H Turner[1], Joshua D Smith[1], Mark J Rieder[1], Koh-ichiro Yoshiura[4], Naomichi Matsumoto[5], Tohru Ohta[6], Norio Niikawa[6], Deborah A Nickerson[1], Michael J Bamshad[1–3] & Jay Shendure[1]

Nature Genet **42**:790–793, 2010

- Seltene Mendel'sche Erkrankung
- Die allermeisten Fälle treten sporadisch auf
- V.a. autosomal dominant

Ng et al.: Kabuki Syndrome (3)

- Wir können nun die statistische Signifikanz schätzen
- Bei einem einzelnen Patienten werden 753 Gene mit Kandidatenmutationen identifiziert
- unser $p = \frac{753}{20.000}$
- Die Wahrscheinlichkeit, dass wir bei genau 7 von 10 Patienten eine Mutation in einem bestimmten Gen (MLL2) sehen, ist daher

```
> p<-753/20000
> dbinom(7,10,p)
[1] 1.146926e-08
```

Ng et al.: Kabuki Syndrome (3)

- Um die statistische Signifikanz zu berechnen, müssen wir die Wahrscheinlichkeit berechnen, dass wir ein mindestens so extremes Ergebnis beobachten. Wir führen zudem eine Bonferroni-Korrektur durch (20.000 Gene: 20.000 Tests!)

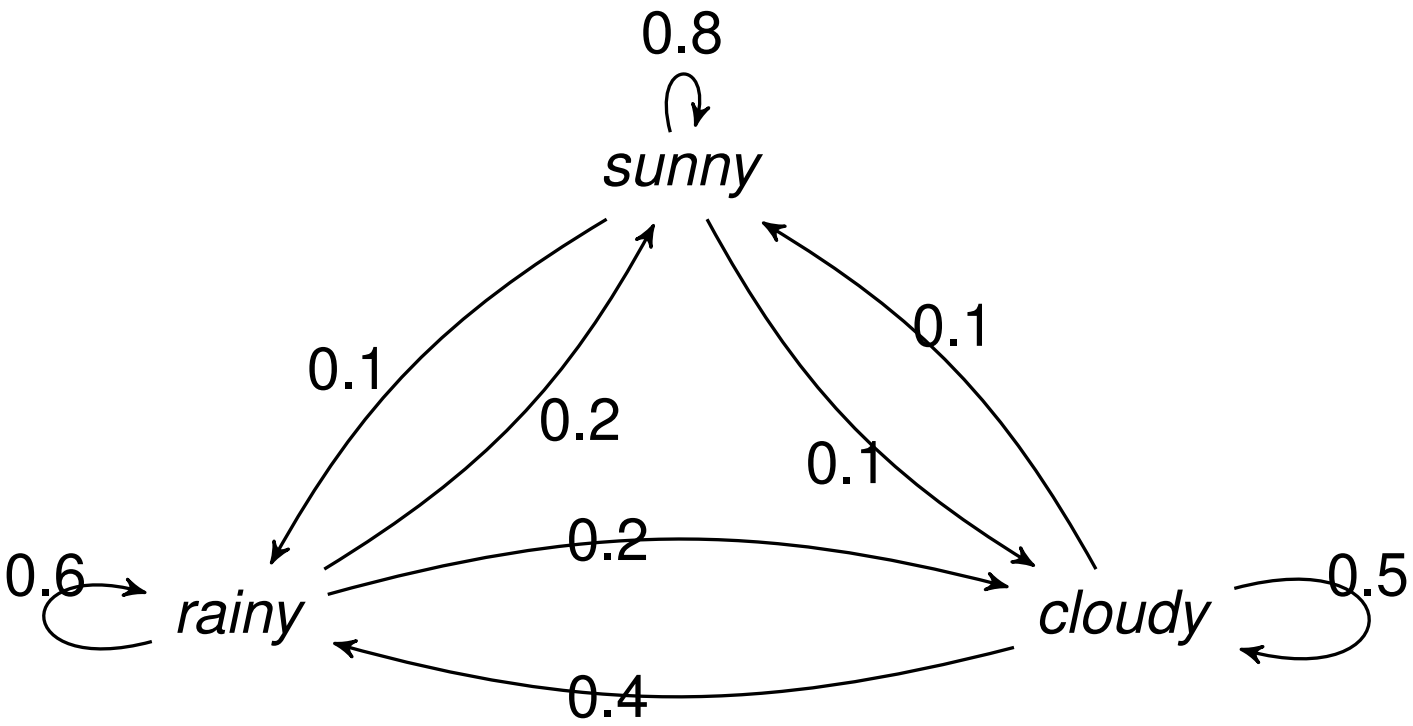
```
> sum(dbinom(7:10,10,p))*20000
[1] 0.0002327798
```

- d.h., wir erhalten einen korrigierten *P*-Wert von 0.0002.
- Es konnte in der Folge von Ng et al. auch bewiesen werden, dass MLL2 das Krankheitsgen für Kabuki-Syndrom ist

Familien-basierte Identifikation von Krankheitsgenen durch NGS

- Die oben vorgestellte Methode funktioniert nur dann, wenn mehrere Patienten mit derselben Krankheit untersucht werden können, was bei seltenen genetischen Krankheiten häufig unmöglich ist
- Im Folgenden wird eine Methode vorgestellt, die für die Untersuchung von einer einzelnen Familie mit einer autosomal rezessiven Erkrankung geeignet ist

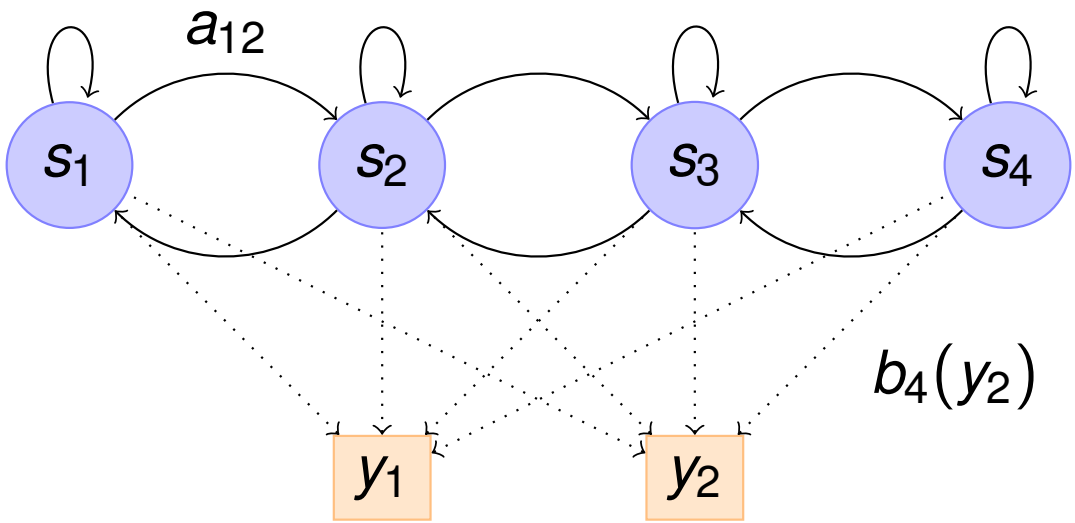
Markow-Kette



- P(SSSSCCCCRRRR)=$0.8^3 \times 0.1 \times 0.5^3 \times 0.4 \times 0.6^3 = 5.5 \times 10^{-4}$
- P(SCRSCRSCRSCR)=$0.1 \times 0.4 \times 0.2 \times 0.1 \times 0.4 \times 0.2 \times 0.1 \times 0.4 \times 0.2 \times 0.1 \times 0.4 \times 0.2 = 4.1 \times 10^{-9}$

Markov-Kette vs. Hidden Markov Model

- Bei einer Markov-Kette können wir die **Zustände** (states) direkt beobachten (e.g., Sunny, Cloudy, and Rainy).
- Bei einem *hidden* Markov model (HMM) können wir die **Zustände** nicht beobachten, sondern lediglich die Emissionen der Zustände



Bayes-Theorem

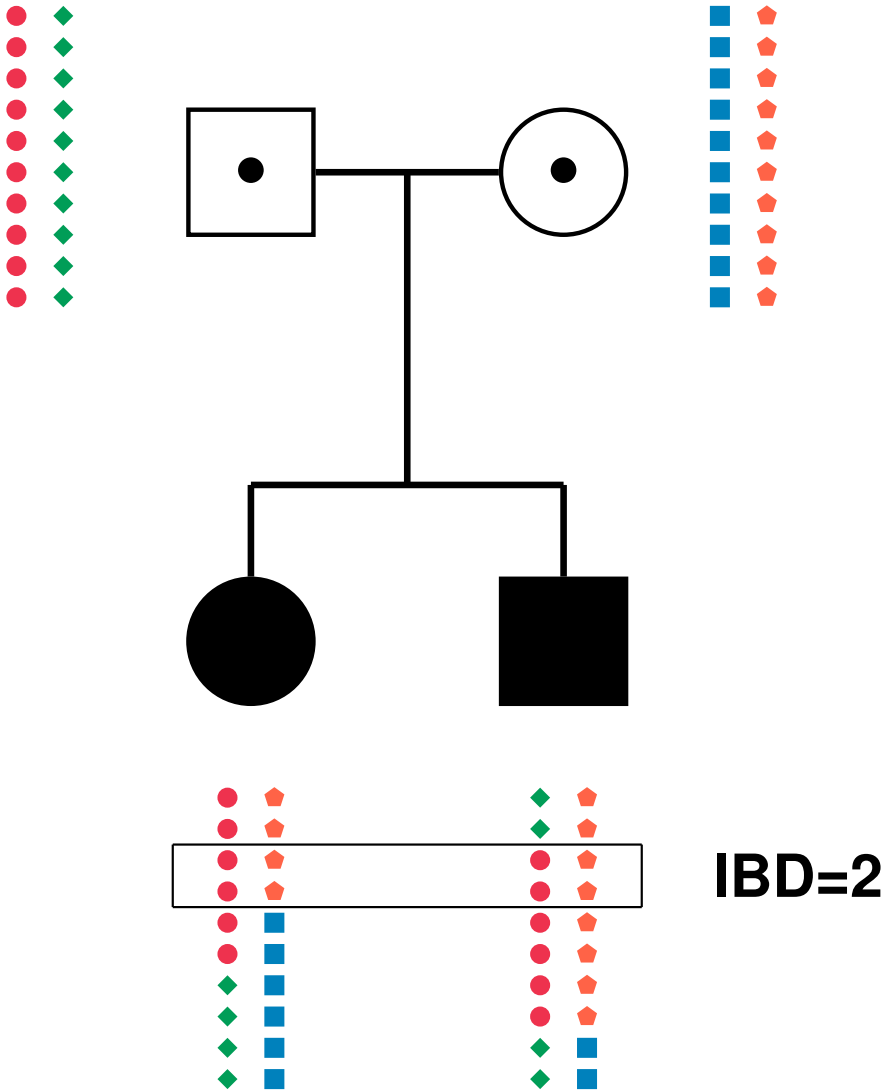
- Ein HMM ist ein Bayes'sches Netzwerk für sequentielle Daten
- Mit dem Bayes-Theorem können wir auf die wahrscheinlichste Reinhenfolge der verborgenen Zustände schließen (**M**) gegeben die beobachteten Daten (**D**)

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}. \tag{8}$$

- Unser Model wird die Reihenfolge von *identical by descent* (IBD) und nicht-IBD Zustände entlang der Chromosome modellieren. Dabei sind die Emissionen die Basenzuweisungen (base calls)

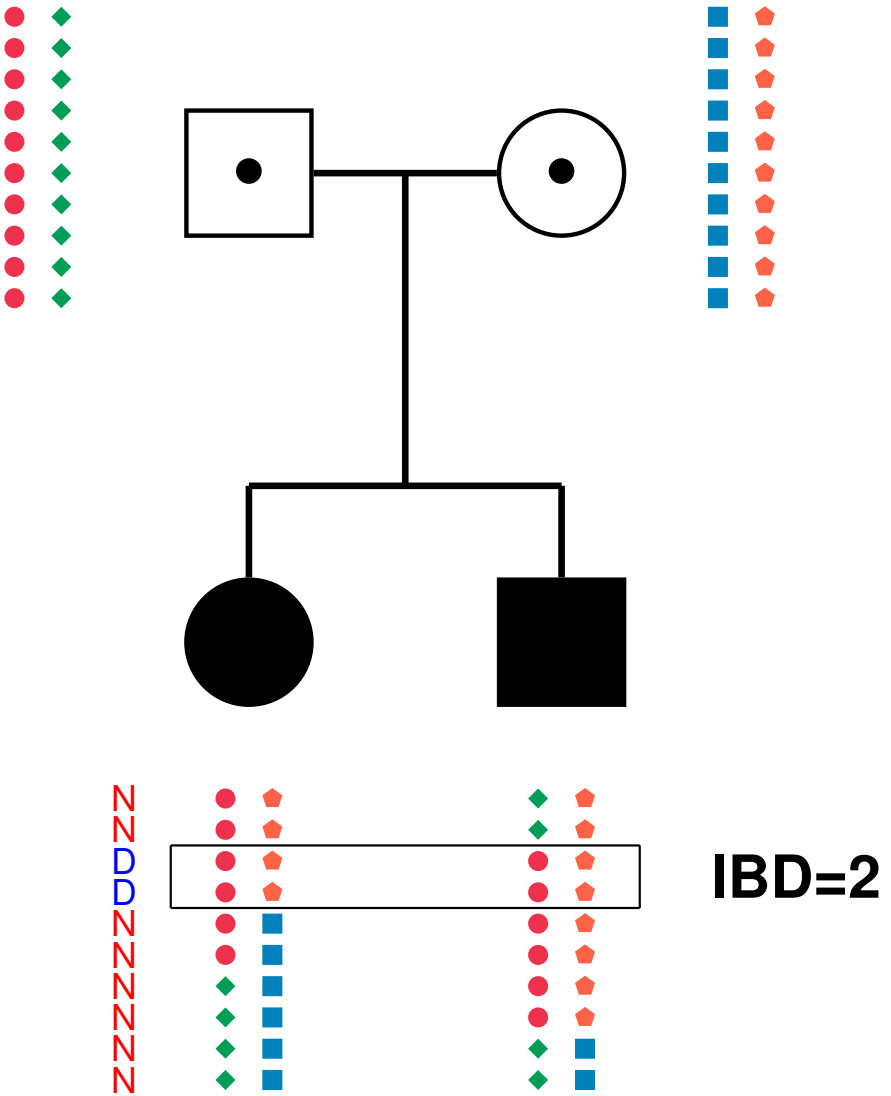
IBD=2

- IBD=2: Identische mütterliche und väterliche Haplotypen
- Bei autosomal rezessiven Erkrankungen muss das Krankheitsgen in einem IBD=2 Bereich gelegen sein
- Weitere IBD=2 Bereiche können auch per Zufall vorkommen

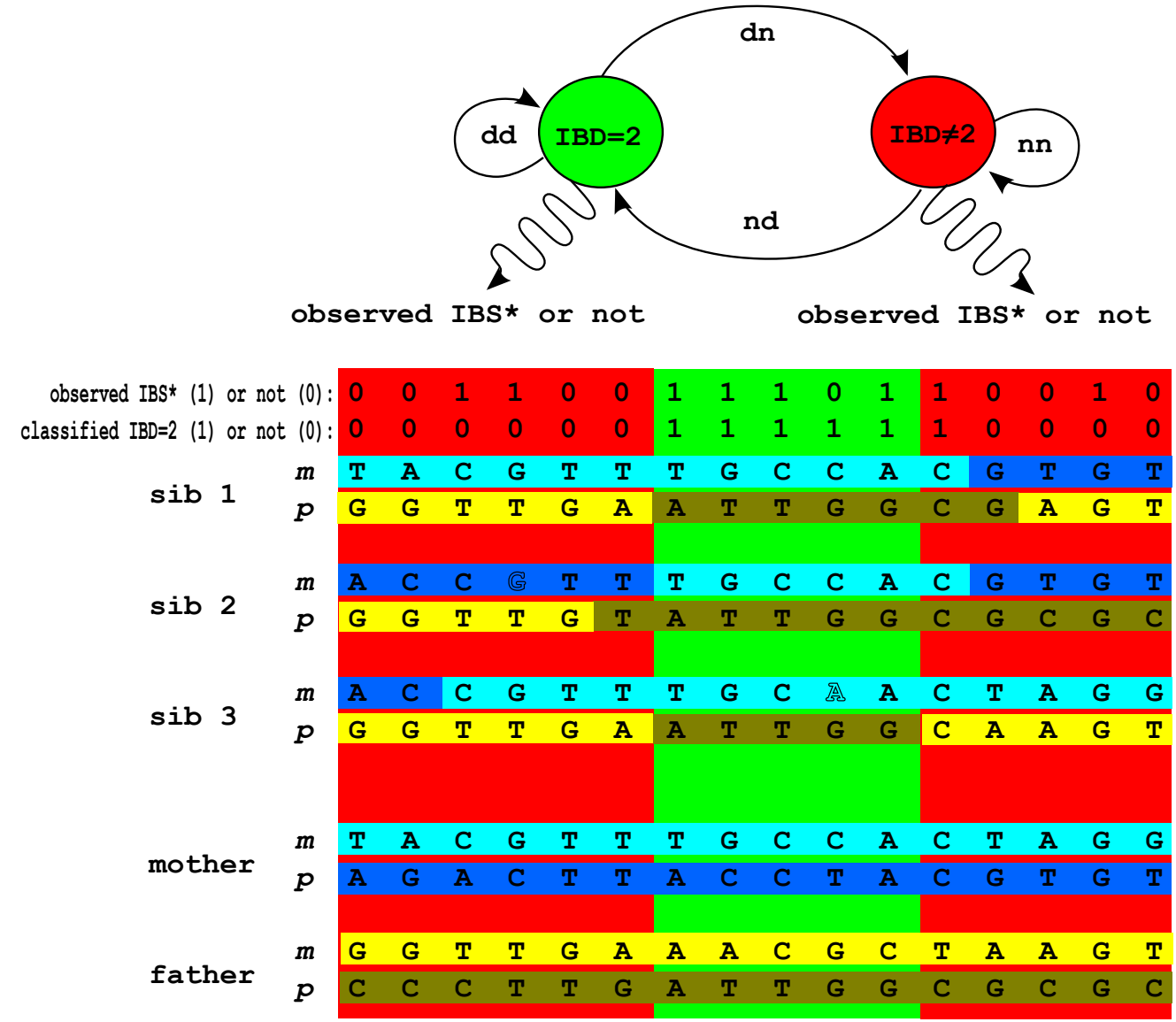


IBD=2

- Die **unbeobachteten Zustände** der chromosomalen Regionen: IBD=2 (D) oder nicht IBD=2 (N)
- Die **Transitionen** zwischen Zuständenhängen von Recombinationen bei einem oder mehreren Geschwistern ab
- Die **Emissionen**: Alle betroffenen Geschwister haben dieselben homozygoten bzw. heterozygoten Varianten (IBS*) oder nicht.



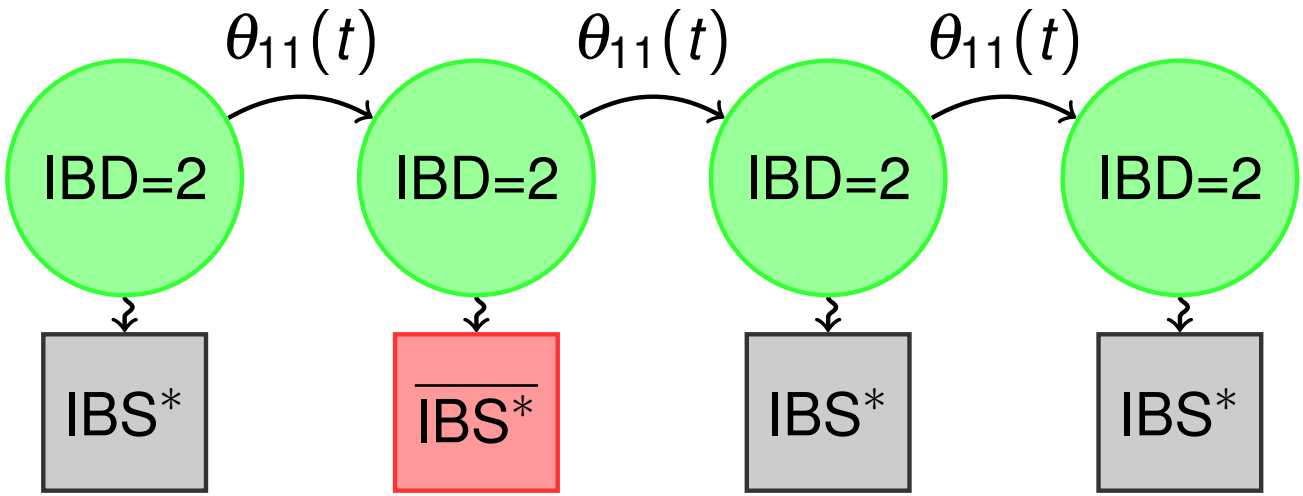
HMM



- Hidden Markov Model mit IBD=2 und IBD≠2 Zuständen

Hidden Markov Model

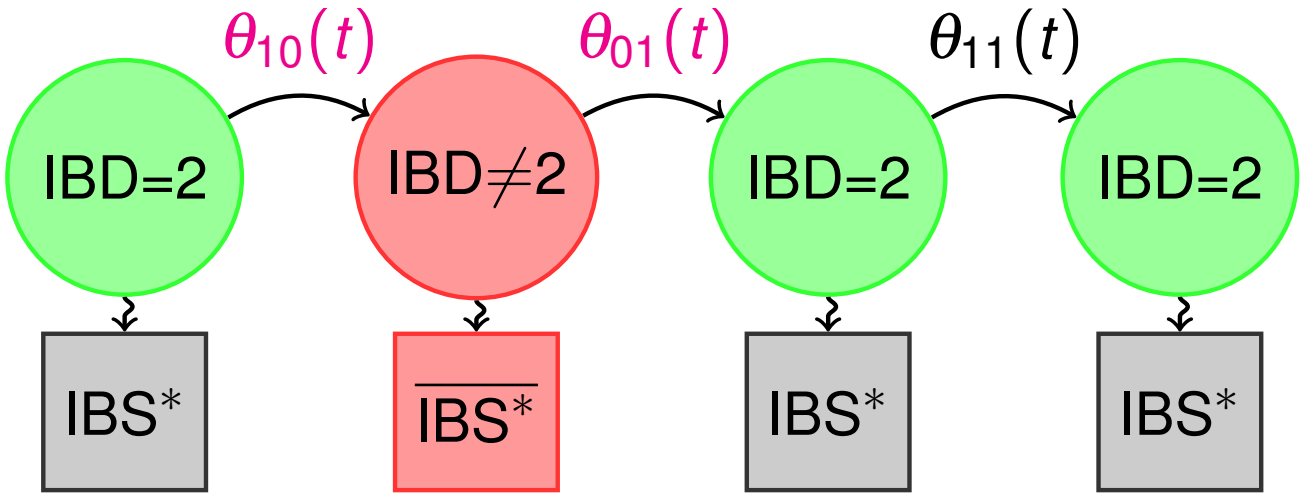
- Unbeobachtete Zustände geben beobachtbare Ausgabesymbole (“Tokens”) aus



- Alle Transitions und Emissionen haben eine verhältnismäßig hohe Wahrscheinlichkeit außer einer IBD=2 $\rightsquigarrow$ $\neg$IBS* Emission (base call error[5] ca. 5%)

[5] engl: Basenzuweisungsfehler

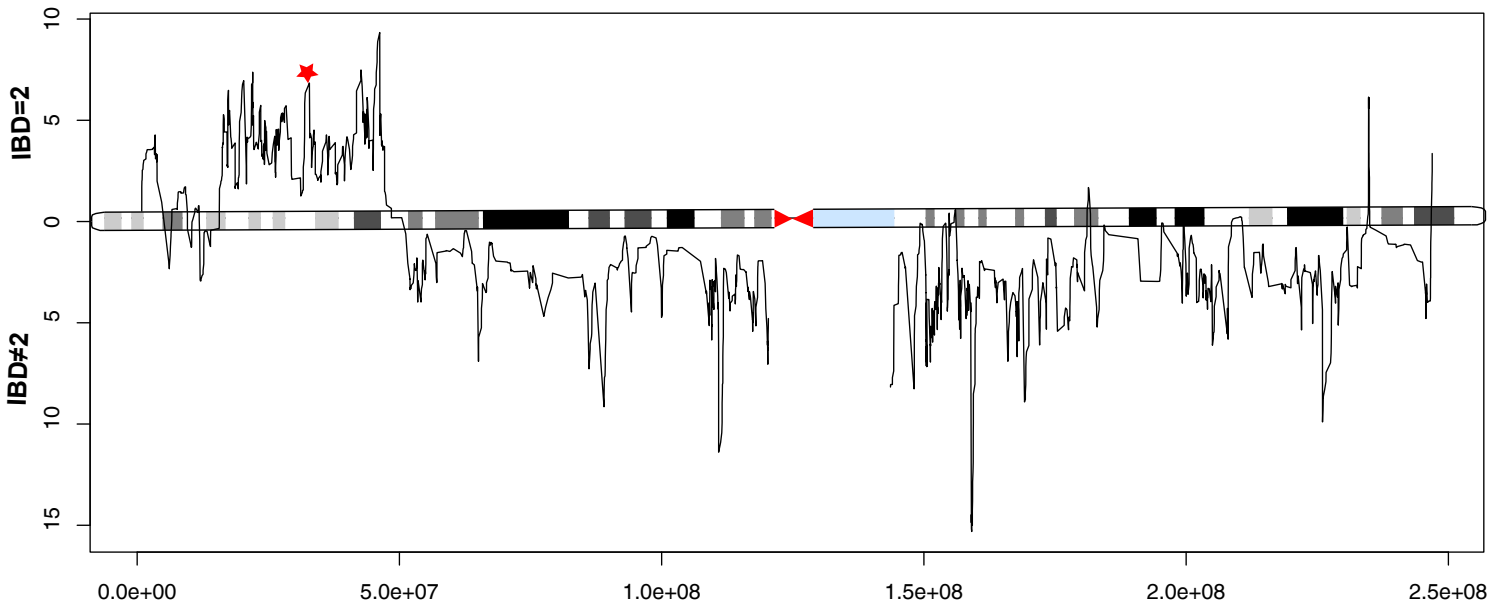
Hidden Markov Model



- Alle Transitions und Emissionen haben eine verhältnismäßig hohe Wahrscheinlichkeit außer den zwei Transitionen $\theta_{10}(t)$ und $\theta_{01}(t)$
- Dies ist sehr unwahrscheinlich: Zwei Rekombinationen innerhalb einer kurzen Entfernung, welche jedoch ohne Basenzuweisungsfehler die ¬IBS*-Beobachtung “erklären”

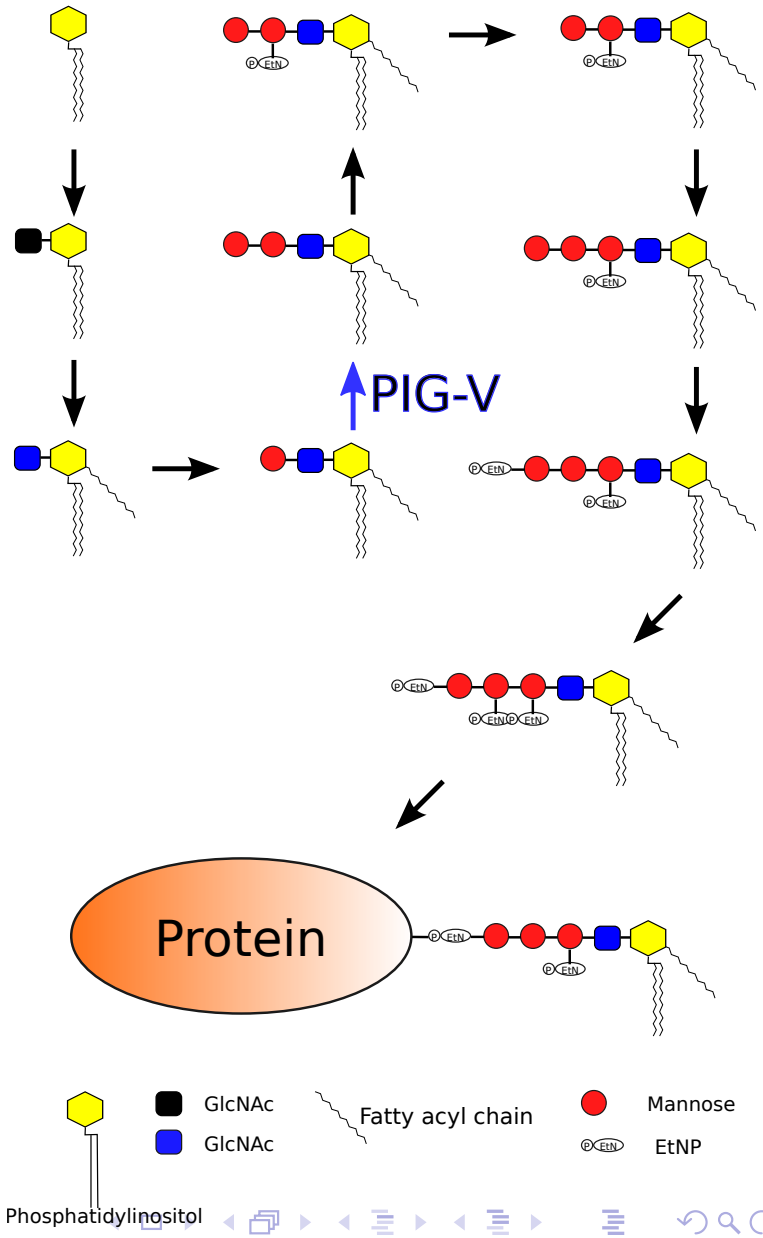
Identifikation von IBD=2 Regionen in einer HPMR

- Das HMM wird mit dem Backward/Forward Algorithmus dekodiert: Somit werden die *A posteriori* Wahrscheinlichkeiten für jede Variante berechnet, vom IBD=2 Zustand ausgegeben worden zu sein.
- Die *A posteriori* Wahrscheinlichkeiten für IBD=2 vs. IBD$\neq$2 kann geplottet werden
- $lod_t = \log_{10} \frac{P(X_t=1|Y=y)}{P(X_t=0|Y=y)}$



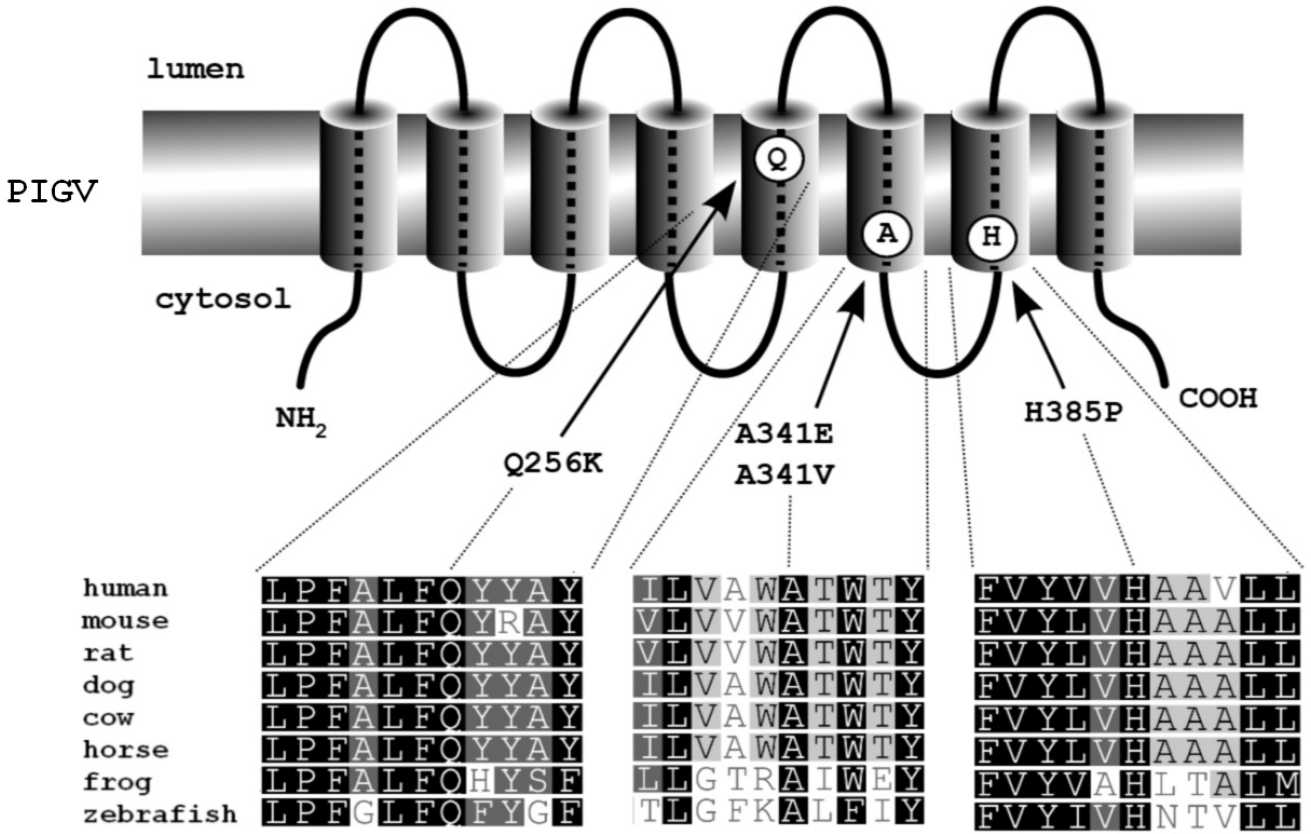
Glycosylphosphatidylinositol (GPI) Pathway

- PIGV, ein Enzym im GPI-Anker-Biosynthese-Pathway, war unter den Genen im IBD=2 Bereich
- *PIGV* codiert für die zweite Mannosyltransferase im GPI-Anker-Biosynthese-Pathway
- $>$ 100 Proteine werden durch einen GPI-Anker am C-Terminus modifiziert



PIGV Mutations Causes HPMR Syndrome

- Homozygote und heterozygote Mutationen sind in drei weiteren Familien nachgewiesen worden



zum Schluss

- Email: peter.robinson@charite.de

weiterführende Literatur

- Gilissen C et al (2012) Disease gene identification strategies for exome sequencing. *Eur J Hum Genet* **20**:490-7.
- Zhi D, Chen R (2012) Statistical guidance for experimental design and data analysis of mutation detection in rare monogenic mendelian diseases by exome sequencing. *PLoS One* **7**:e31358.
- Krawitz et al, (2010) Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nat Genet* **42**:827–829.