

Next-Generation Sequenzierung in der Genetik am Beispiel der Harnstoffzyklusstörungen

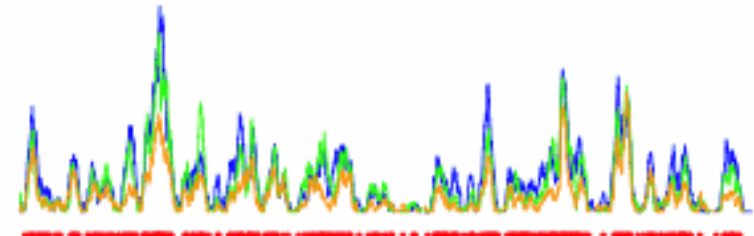
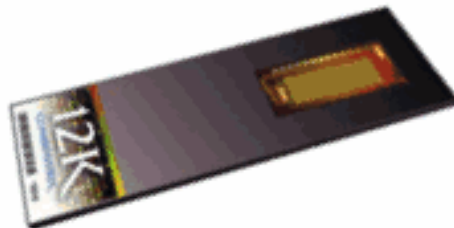
Carlo Largiadèr

Universitätsinstitut für Klinische Chemie UKC
Molekulare Diagnostik DOLS

3. März, 2011

 **INSELSPITAL**
UNIVERSITÄTSSPITAL BERN
HOPITAL UNIVERSITAIRE DE BERNE
BERN UNIVERSITY HOSPITAL

u^b
UNIVERSITÄT
BERN



Inhalt

Aufzeigen der Möglichkeiten und Probleme bei der Etablierung der NGS-Technik im diagnostischen Labor anhand eines Beispiels

Clinical Chemistry 57:1
102–111 (2011)

Molecular Diagnostics and Genetics

Sequence Capture and Next-Generation Resequencing of Multiple Tagged Nucleic Acid Samples for Mutation Screening of Urea Cycle Disorders

Ursula Amstutz,^{1,2} Gisela Andrey-Zürcher,¹ Dominic Suci,³ Rolf Jaggi,⁴
Johannes Häberle,⁵ and Carlo R. Largiadèr^{1*}

=> Bei welchen Schritten hatten wir den höchsten Aufwand?



Grundlegende Schritte in der Entwicklung eines NGS-basierten Tests

- 1) Vorbereitung der Ziel-DNA (Multiplex-PCR, DNA-Barcoding-Strategien oder Sequenzanreicherung) und evtl. Erarbeitung von Referenzdaten für die Validierung des Tests
- 2) Probenvorbereitung für die Sequenzierung & Sequenzierung
- 3) Erste Grundausswertung der Sequenzierdaten
- 4) Spezifische Datenauswertung (Validierung)



Next-Generation Sequenzierung (NGS)

„High-throughput sequencing“:

→ **Gigabasen** von Sequenzdaten können innerhalb weniger Tagen generiert werden

Die am verbreitetsten Plattformen:

- 454 sequencing
- Illumina
- SOLiD

	454 FLX	Illumina	SOLiD
Read length	400 bp	35-100 bp	75 bp
Reads/run	1 million	1-2 billion	2.8 billion
Output	400 Mb	26-200 Gb	300 Gb
Run time	10 hours	2-8 days	1-7 days

weniger aber
längere *Reads*

viele aber
kurze *Reads*



Sequenzierkosten

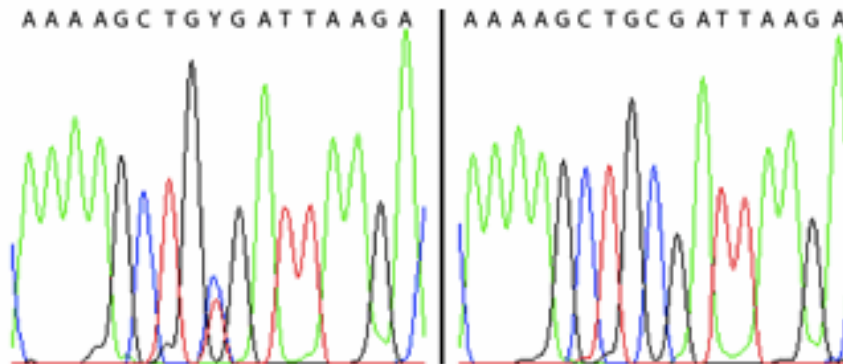


Baker, Nature Methods 2010



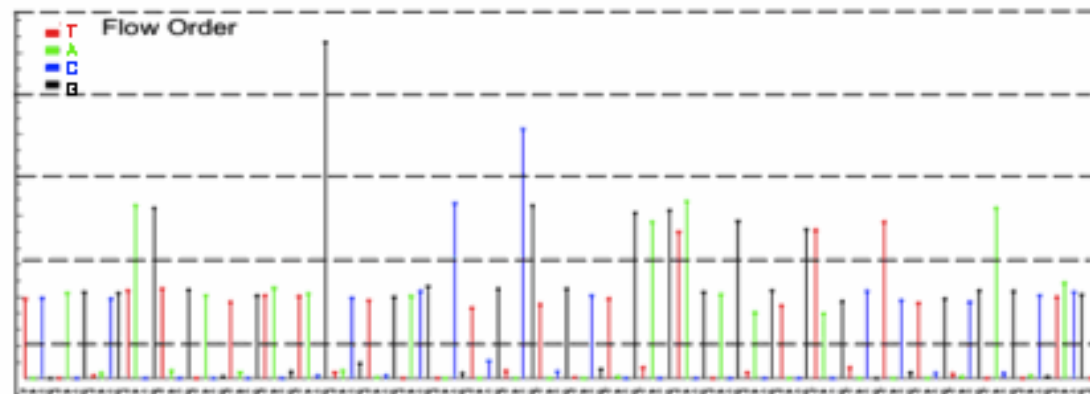
NGS vs. Sanger-Sequenzierung

Sanger sequencing



1 Probe → Pool von Fragmenten → 1 Sequenz

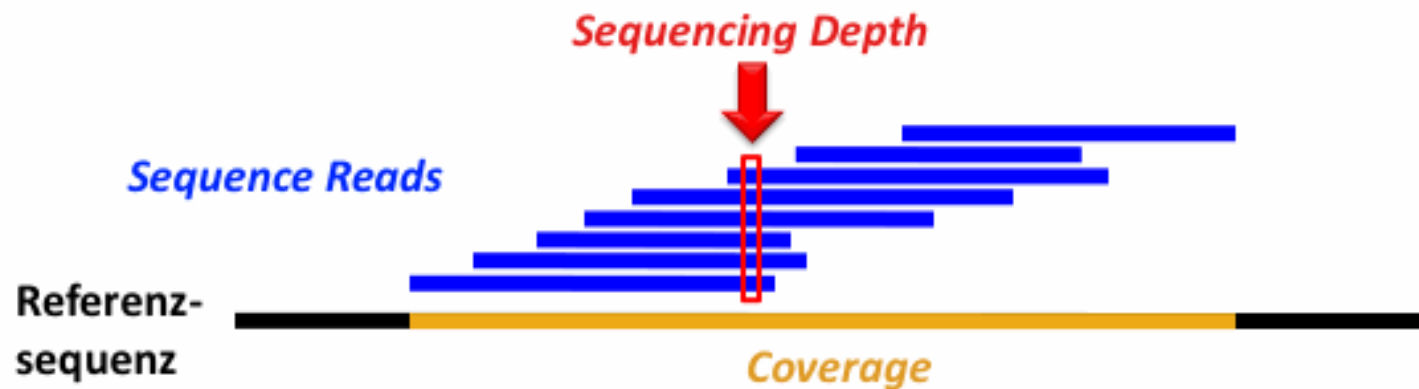
next-generation sequencing



1 DNA Fragment → 1 *Bead* → 1 Sequenz-Read
1 Probe → viele Fragmente → viele Sequenz-Reads



Eigenschaften von NGS-Daten



- **Sequence Read**: Sequenz von einem klonal-amplifizierten Fragment stammend
- **Coverage**: Anteil der Zielregion welche durch ≥ 1 *Sequence Read* abgedeckt wird
- **Sequencing Depth/Read Depth**: Die Häufigkeit mit der eine bestimmte Position sequenziert wurde

1 DNA Fragment \rightarrow 1 Bead \rightarrow 1 Sequenz-Read

\rightarrow Genügende *Sequencing Depth* zur Detektion von Varianten benötigt



Berechnungsbeispiel für die Berechnung des Sequenzieraufwandes (Kosten pro bp Zielsequenz)

System produziert 2.8 Mb Sequenzdaten pro Run

Mittlere *Read*-Länge von 400 bp

70'000 *Reads* erwartet (entspricht ca. Roche GS Junior System)

Für eine zuverlässige Heterozygoten-Detektion brauchen wir theoretisch (Binomial-Verteilung) eine *Sequencing Depth* von > 20-fach

(=> Wahrscheinlichkeit für eine Frequenz eines der Allele mit einer Frequenz von >10% zu beobachten: 99,97%)

⇒ Max. kann eine Zielsequenz 1,4 MB oder z.B. 7 Proben mit einer Zielsequenz von 200 Kb mit einer genügenden *Coverage* sequenziert werden.

⇒ In der Realität braucht es deutlich mehr Reads!



Aufwand und Grösse der Zielsequenz

Zusätzliche Kosten

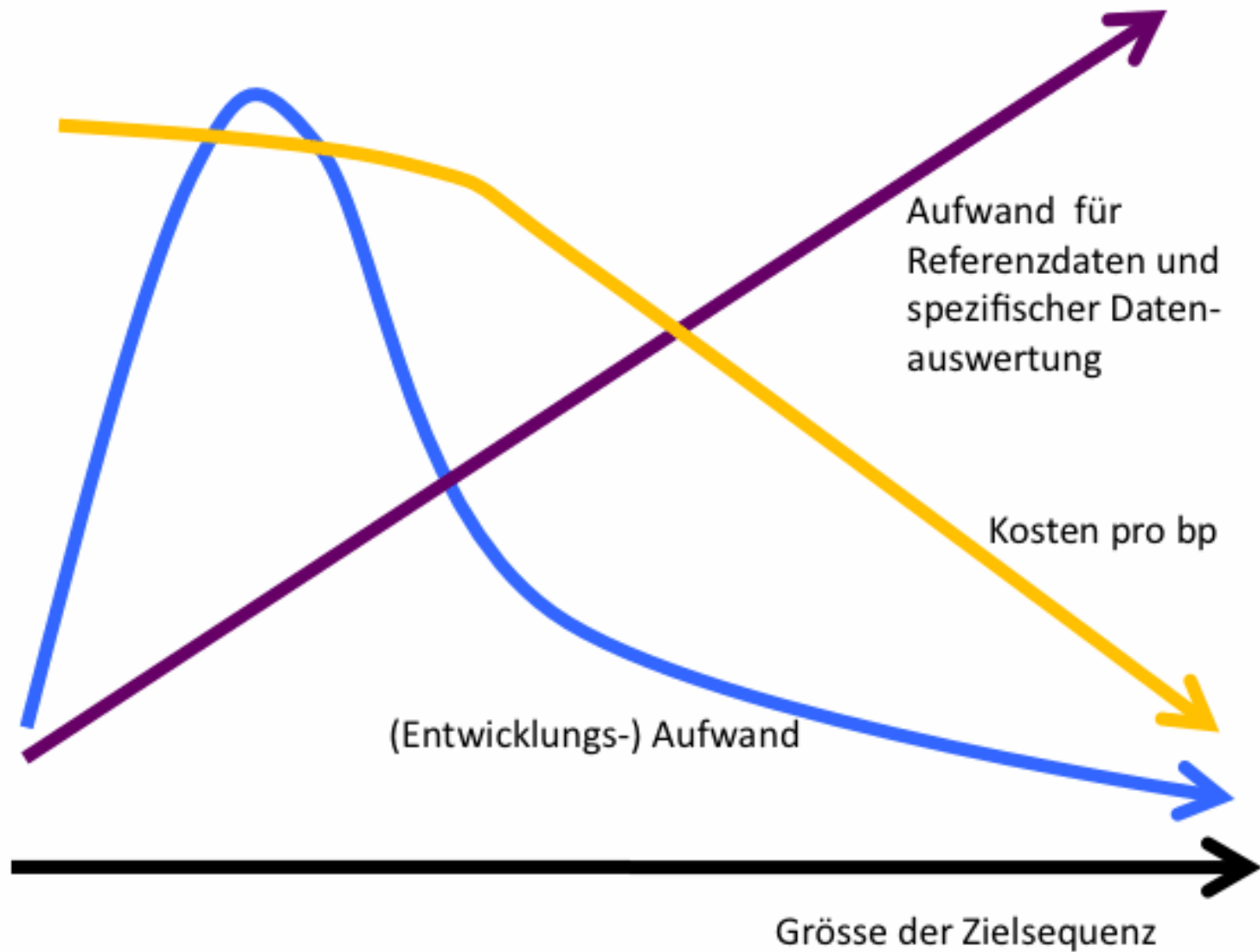
(Barcodes; Anreicherungsverfahren;
Multiplex-PCRs, mehrere *Sequencing libraries* etc.)

(Entwicklungs-) Aufwand

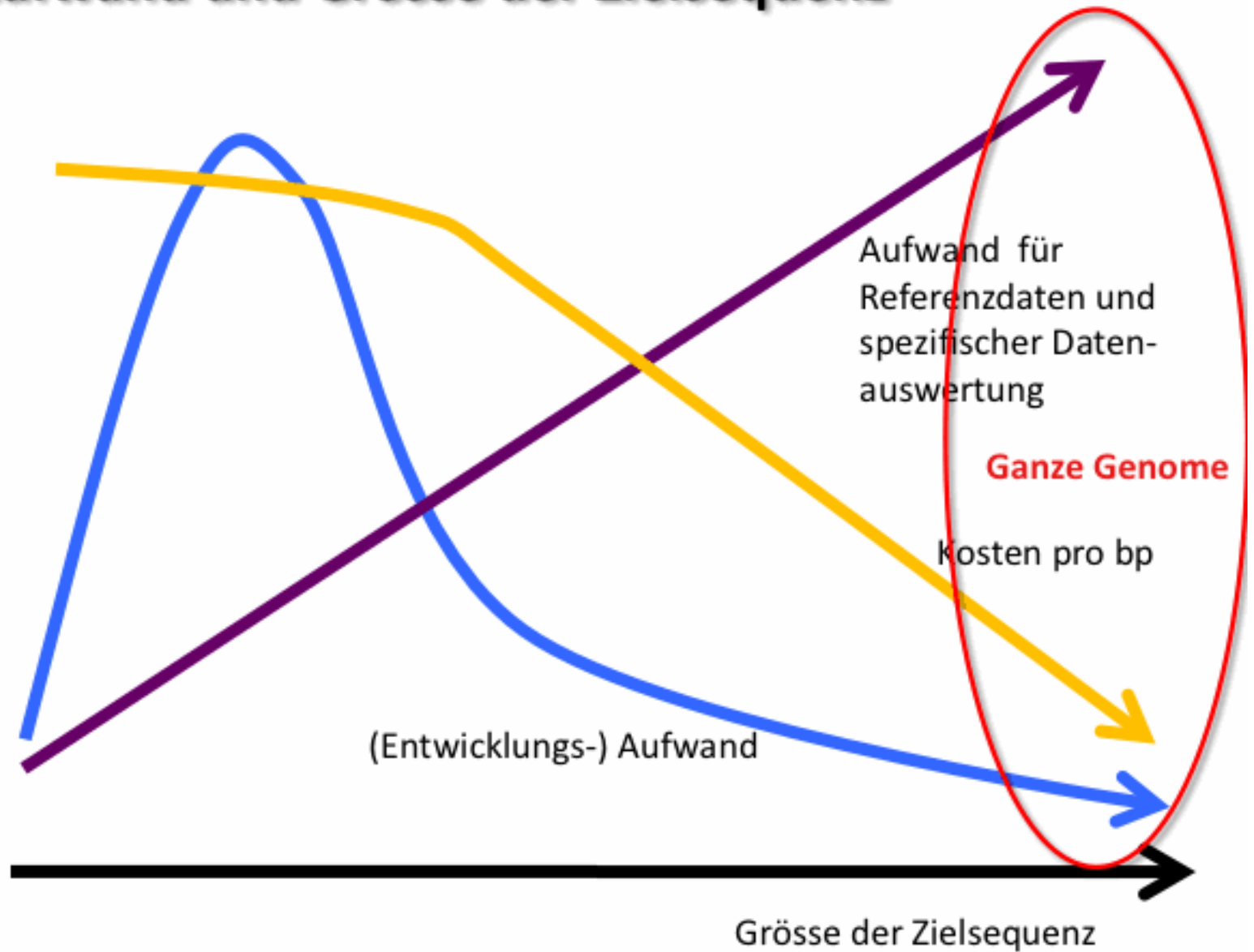
Grösse der Zielsequenz



Aufwand und Grösse der Zielsequenz



Aufwand und Grösse der Zielsequenz



Diagnostik mit ganzen Genomen?

Nature. 2008 Apr 17;452(7189):872-6.

The complete genome of an individual by massively parallel DNA sequencing.

Wheeler et al.

.....Comparison of the sequence to the reference genome led to the identification **of 3.3 million single nucleotide polymorphisms**, of which **10,654 cause amino-acid substitution within the coding sequence**. In addition, we accurately identified small-scale (2-40,000 base pair (bp)) insertion and deletion polymorphism as well as copy number variation resulting in the large-scale gain and loss of chromosomal segments ranging from 26,000 to 1.5 million base pairs.....

Problem: Was machen wir mit all dieser Information?



Diagnostik mit ganzen Genomen?

ARTICLE

doi:10.1038/nature09534

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

The 1000 Genomes Project aims to provide a deep characterization of human genome sequence variation as a foundation for investigating the relationship between genotype and phenotype. Here we present results of the pilot phase of the project, designed to develop and compare different strategies for genome-wide sequencing with high-throughput platforms. We undertook three projects: low-coverage whole-genome sequencing of 179 individuals from four populations; high-coverage sequencing of two mother-father-child trios; and exon-targeted sequencing of 697 individuals from seven populations. We describe the location, allele frequency and local haplotype structure of approximately 15 million single nucleotide polymorphisms, 1 million short insertions and deletions, and 20,000 structural variants, most of which were previously undescribed. We show that, because we have catalogued the vast majority of common variation, over 95% of the currently accessible variants found in any individual are present in this data set. On average, each person is found to carry approximately 250 to 300 loss-of-function variants in annotated genes and 50 to 100 variants previously implicated in inherited disorders. We demonstrate how these results can be used to inform association and functional studies. From the two trios, we directly estimate the rate of *de novo* germline base substitution mutations to be approximately 10^{-8} per base pair per generation. We explore the data with regard to signatures of natural selection, and identify a marked reduction of genetic variation in the neighbourhood of genes, due to selection at linked sites. These methods and public data will support the next phase of human genetic research.

Geplant : Sequenzierung von 2500 Genomen



Angeborene Harnstoffzyklusdefekte

- Vererbare Stoffwechselerkrankungen; Inzidenz von 1: 8000
- Harnstoffzyklus: **einzig**er Stoffwechselweg um überschüssiges Ammoniak zu metabolisieren
- Enzymdefekte führen zu einer **toxischen** Hyperammonämie

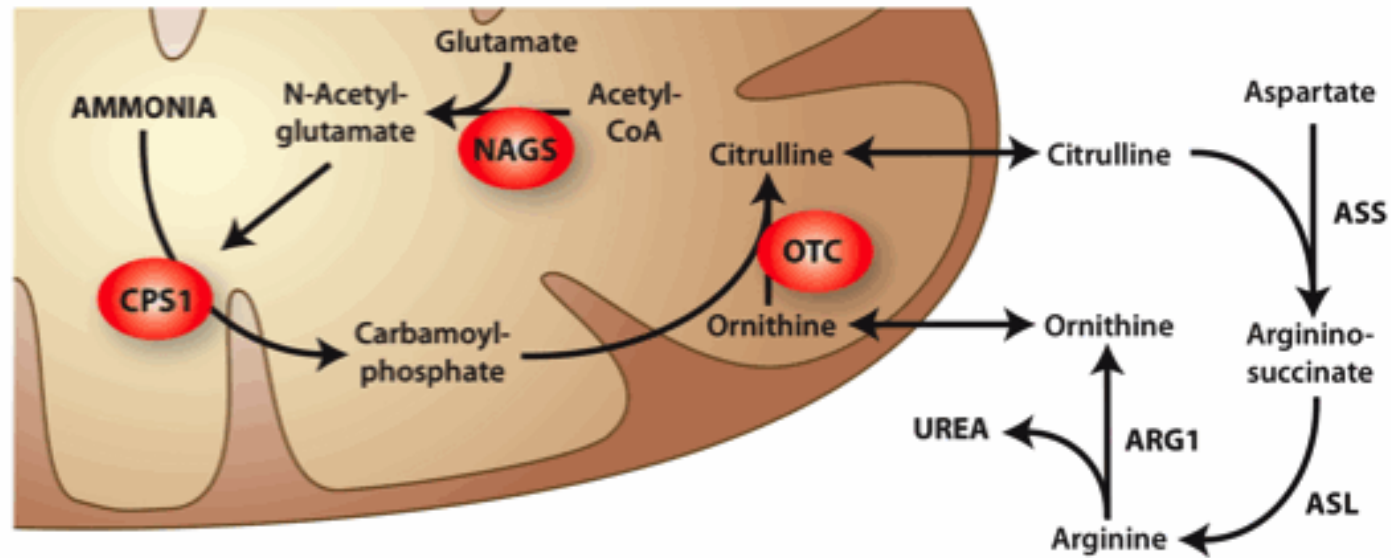


**Krampfanfälle, Enzephalopathie,
Koma**

- Hohe Morbidität und Mortalität
- Diagnose: Messungen von Ammoniak, Aminosäuren, Orotsäure; **genetisches Mutationscreening** zur Diagnosebestätigung



Harnstoffzyklus



6 Enzyme:

CPS1:	Carbamoylphosphat-Synthetase I	} Mitochondrium
NAGS:	N-Acetylglutamat-Synthetase	
OTC:	Ornithin-Transcarbamylase	
ASS:	Argininosuccinat-Synthase	} Cytosol
ASL:	Argininosuccinat-Lyase	
ARG1:	Arginase 1	



Mutationsscreening für OTC-, CPS1- und NAGS-Mangel

→ OTC-Mangel

- häufigste HZS; X-chromosomal
- >340 Mutationen bekannt
- **In 20%** der Fälle **keine Mutation** in den Exons/Spleissstellen beobachtet

→ CPS1-Mangel

- autosomal rezessiv
- grosses Gen: >120 kb; **38 Exons**

→ NAGS-Mangel

- autosomal rezessiv
- **Identische** biochemische Präsentation wie CPS1-Mange



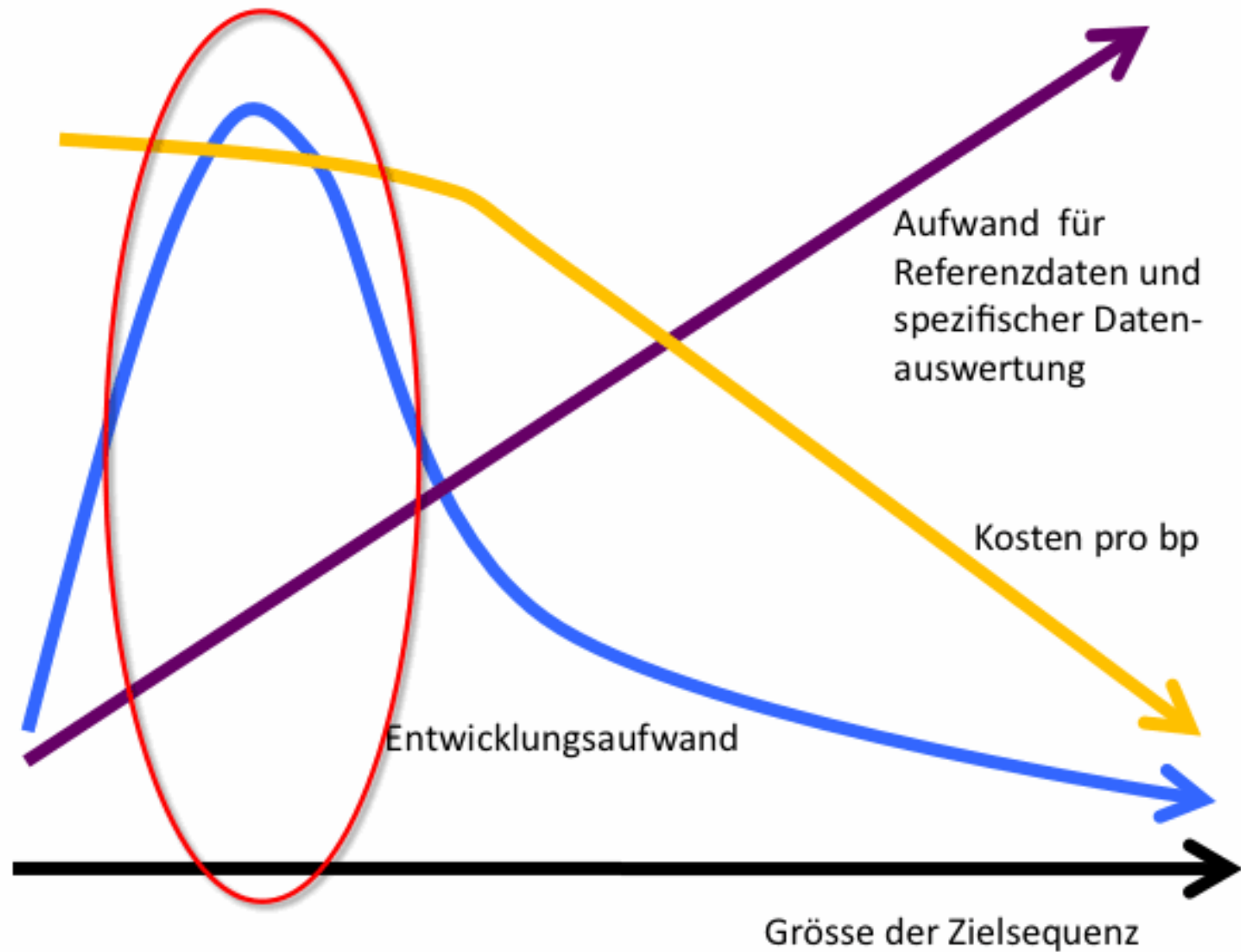
**Simultanes Screening von nicht-kodierenden
Genregionen, vielen Exons und mehre Genen**

Zielregion:

OTC (70Kb), NAGS (5.6Kb), CPS1 (124Kb) = **199 Kb**



Aufwand und Grösse der Zielsequenz



Array-basierende Sequenzanreicherung

Gezielte Anreicherung von spezifischen Regionen in komplexen eukaryotischen Genomen (*Sequence capture*):

Traditioneller Ansatz: **PCR** → **limitiert** bezüglich
Produktlänge und Multiplexing-Grad
(ca. 1000 PCR-Reaktionen für NGS-Sequenzierung
Amplikons in diesem Beispiel notwendig)

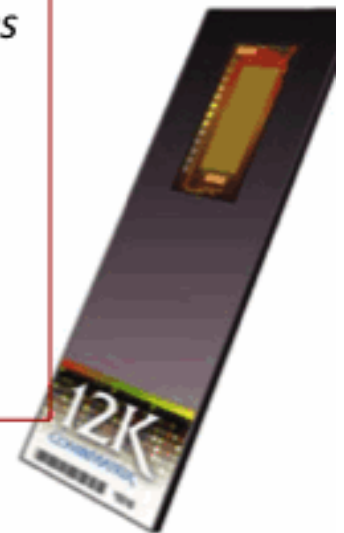


Verwendung von komplementären **Oligonukleotid-Probes**
um Zielregion anzureichern (e.g. DNA-Arrays)



keine Limitierung bezüglich **räumliche Verteilung** der
Zielregionen

(verteilte kurze Segmente & einzelne lange Segmente)



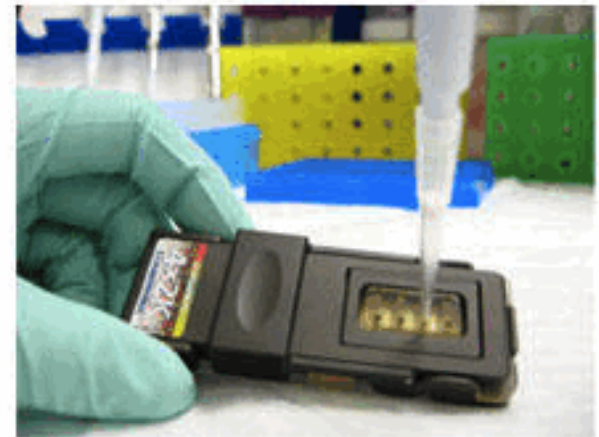
Array-basierende Sequenzanreicherung

1. Schritt: Probe Design

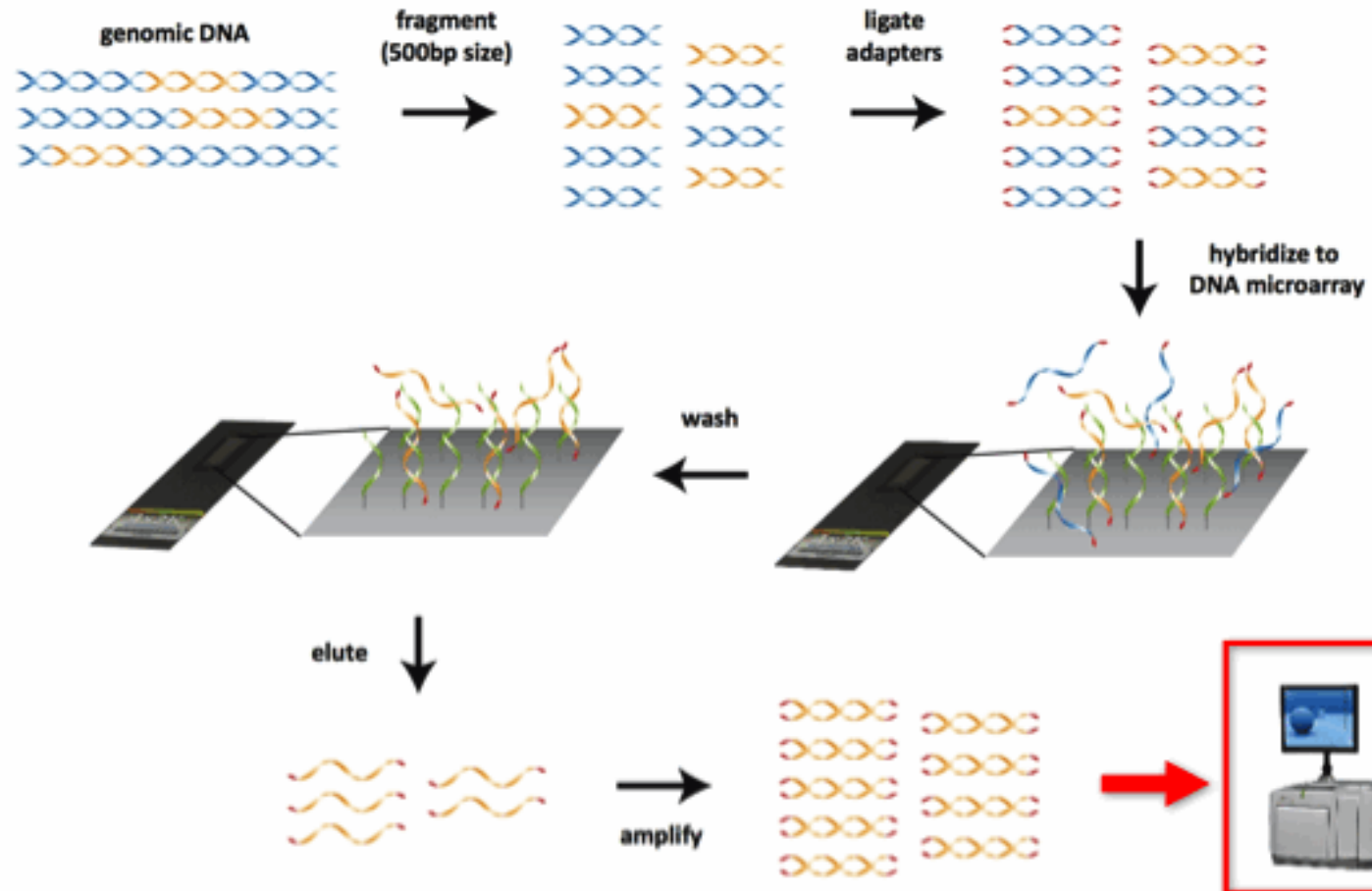
- Aufteilung der Zielregion in 30 bp -Abschnitten
 - Probe-Länge: **50-60 bp** ($T_m = 78^\circ\text{C}$)
 - **Qualitätsscore** (Komplexität, Sekundärstrukturen, GC-Gehalt)
 - **Spezifität**: BLAST –Vergleich mit Humanen Genom
- Die besten *Probes* ausgewählt

→ Tiefe *Tiling Density*: **1 Probe / 91 bp**
(total **2240 Probes**)

→ Sektorierte Arrays: **4 Proben / Chip**



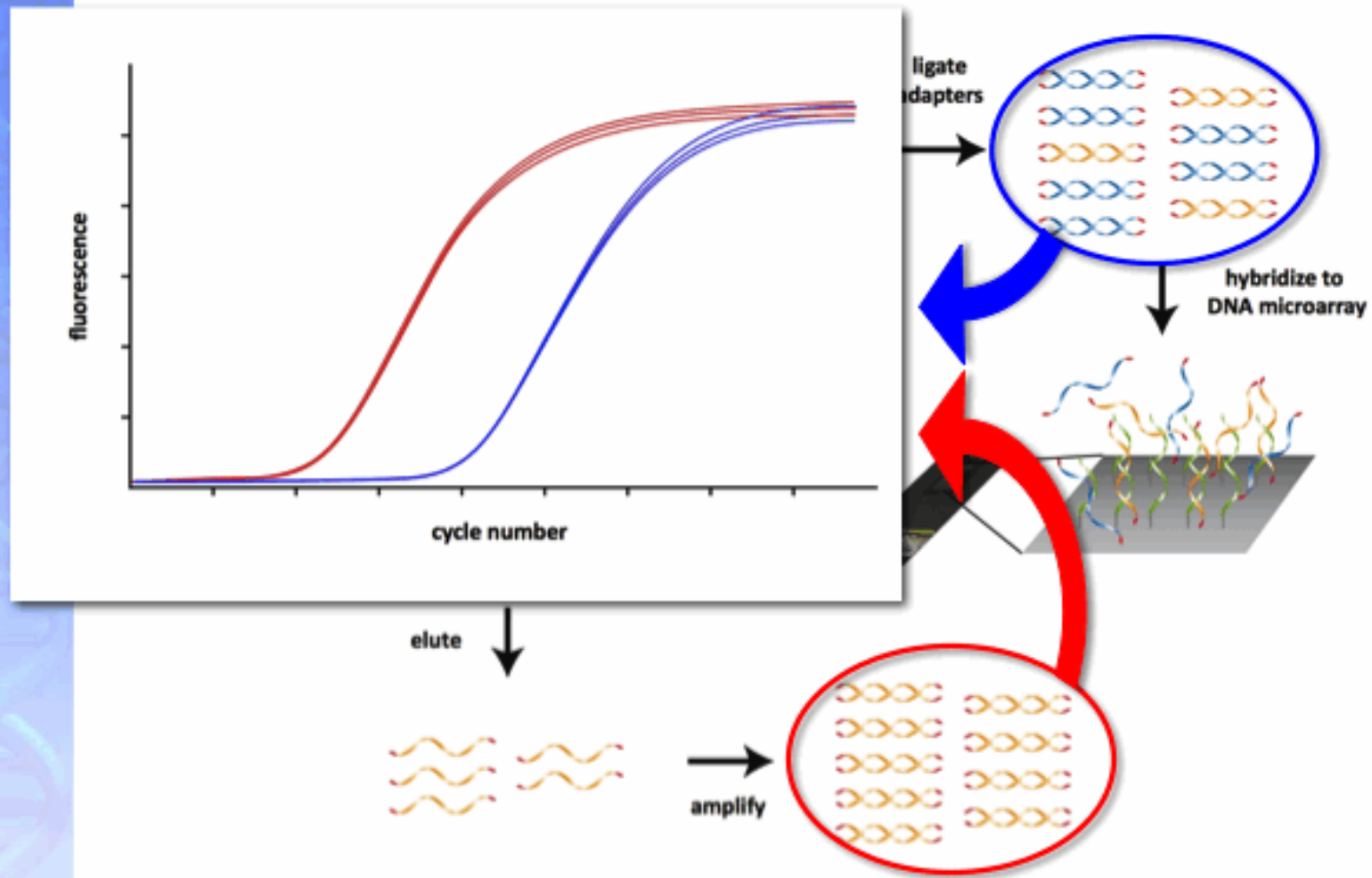
Array-basierende Sequenzanreicherung: Arbeitsablauf



454 sequencing



Messung der Anreicherungseffizienz

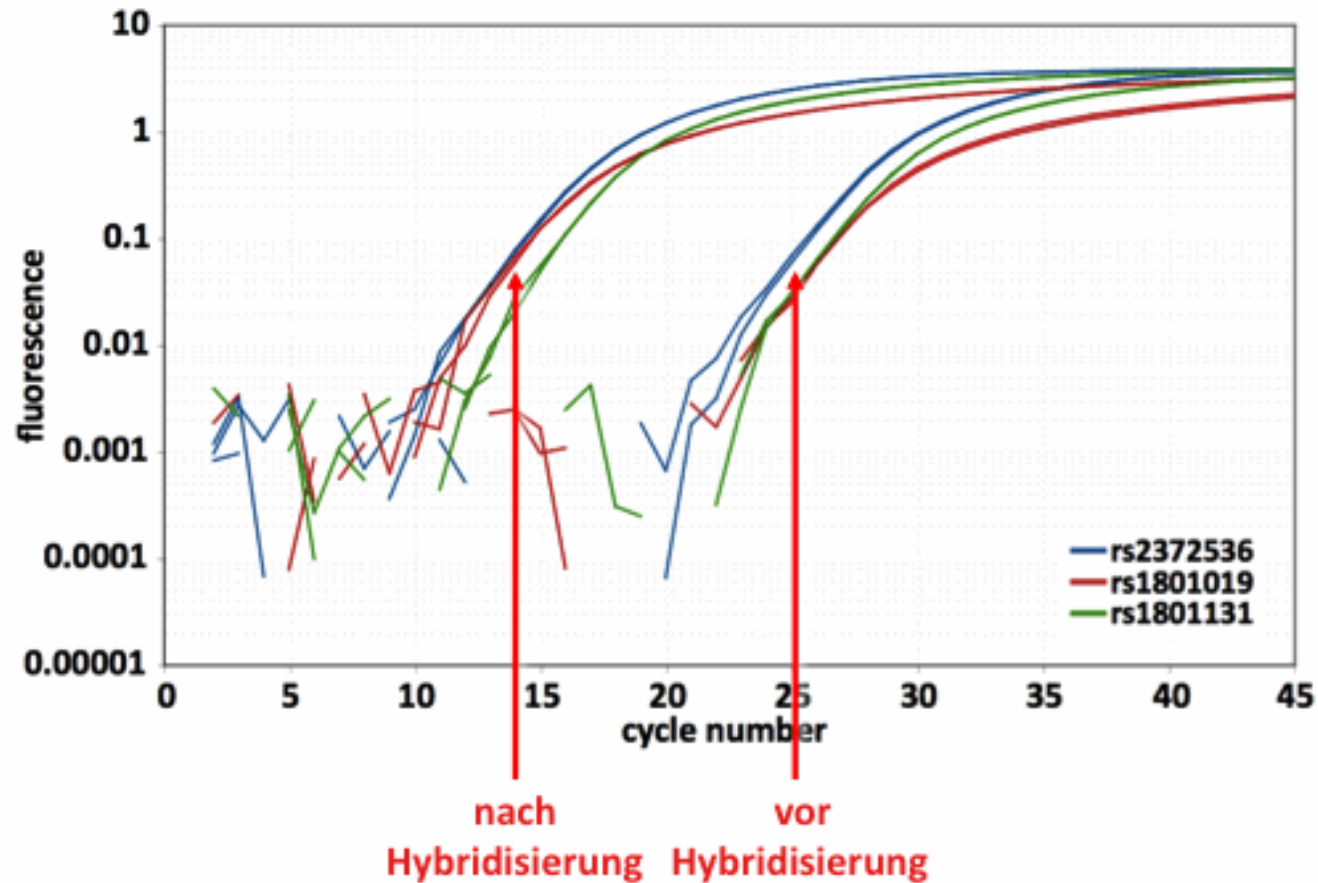


→ Vergleich der Probe **vor** und **nach** der Array-Hybridisierung mittels qPCR



Messung der Anreicherungs-effizienz

Vergleich der Probe vor und nach der Array-Hybridisierung mittels qPCR (3 Loci):



$\Delta Ct \approx 11 \rightarrow \approx 2000\text{-fache}$ Anreicherung



Experimentelles Design HZS-Studie

4 Patientenproben:

- 2 Träger von OTC-Mutationen (OTCA und OTCB)
- 2 Träger von CPS1-Mutationen (CPSA und CPSB)

„**Negative Kontrolle**“: Probe OTCB ohne Array-basierende Sequenzanreicherung (UNC)

	OTCA	OTCB	CPSA	CPSB	UNC
capture	yes	yes	yes	yes	no
barcode	no	yes	yes	yes	yes
library	individual	pooled			
Sequencing	1/8 plate	1/2 plate			

(Roche GS 454 FLX)



Coverage und Sequencing Depth

- Ca. 480'000 *Sequence Reads*
- **91-95%** der Zielregion abgedeckt (>180 Kb)
- *Sequencing Depth* von **9-15**-fach (Median)
- **>3 Mb** Sequenzdaten innerhalb der Zielregion pro Probe

	bp on target	coverage bp	coverage %	sequencing depth median	sequencing depth IQR
CPSA	3424114	187609	94.1	11	(4-23)
CPSB	3089499	187193	93.8	10	(4-21)
OTCB	3868424	189895	95.2	15	(6-26)
OTCA	3206460	181144	90.8	9	(3-23)

IQR = interquartile range

=> Nicht alle Regionen gleich gut anreichern.

=> Total ca. 500'000 *Sequence Reads* werden für eine Probe benötigt, um 75% der Zielregion mit einer 20-fachen *Coverage* zu sequenzieren.



Sequenzeigenschaften und Anreicherungseffizienz

	CPSA		CPSA		OTCB		OTCA	
	coverage	depth	coverage	depth	coverage	depth	coverage	depth
OTC	90.3%	7	93.1%	10	93.4%	13	92.7%	11
NAGS	90.8%	3	94.5%	5	81.5%	3	94.6%	3
CPS	96.4%	15	94.2%	11	96.8%	16	89.6%	8

Ähnliche *Coverage* aber **tiefer** *Sequencing Depth* (Median) für das **NAGS -Gen**

→ Vergleich der Eigenschaften der *Probes* zwischen den Zielsequenzen

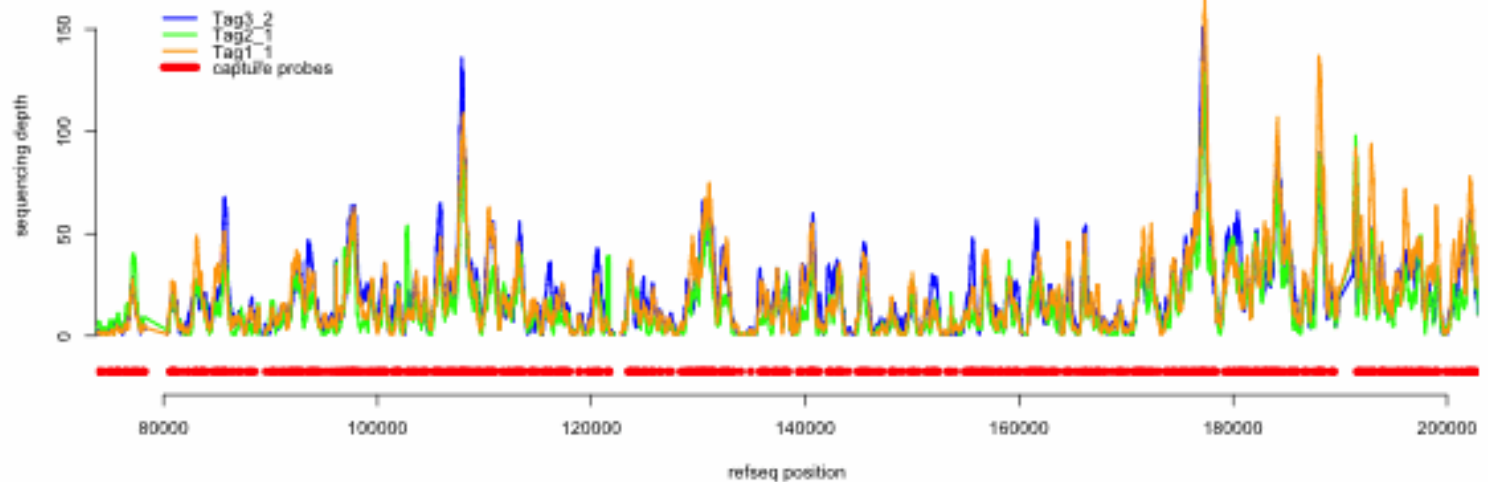
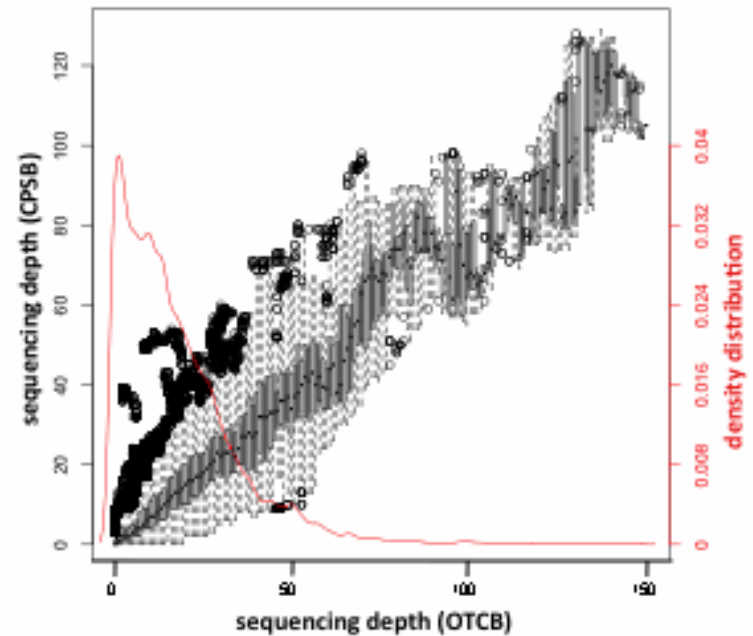
	tiling density (bp per probe)	average length (bp)	T _m (°C)	GC-content (%)
OTC	91	57	77.4	42.6
NAGS	98	51	81.1	54.4
CPS	90	58	76.8	41.0

höheres T_m, höherer GC-Gehalt, kürzere Länge, tiefere Tiling Density
⇒ **NAGS GC-reiche Sequenz**



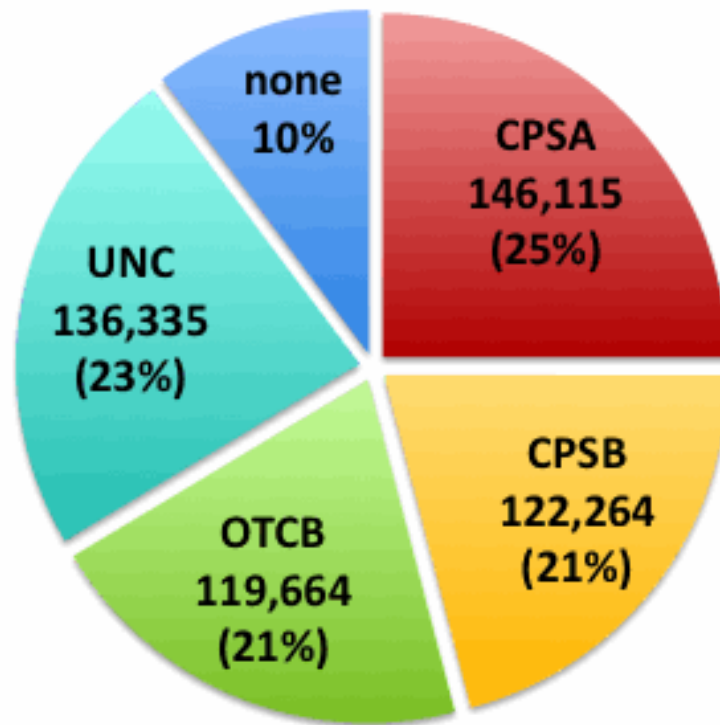
Reproduzierbarkeit

- **Starke Korrelation** zwischen der Sequencing Depth between samples
- Spearman's rank correlation coefficient: **0.84-0.86**
- **>80%** der Zielregion mit einer *Depth* ≥ 10 in der Probe CPSB weist die **gleiche oder höhere** *Depth* in den Proben CPSA und OTCB auf



Homogenität des Multiplex-Ansatzes

90% der *Sequence Reads* können aufgrund des DNA-Barcodes eindeutig einer Probe zugeordnet werden



DNA-Barcoding:

½ Platte (FLX) für 4 Proben
120'000-146'000 Reads/Probe

Ohne Barcode (OTCA):

½ Platte (FLX) für 1 Probe
90'000 Reads/Probe

Homogenität wird beeinflusst die Qualität der einzelnen Proben und der Genauigkeit beim Multiplexing

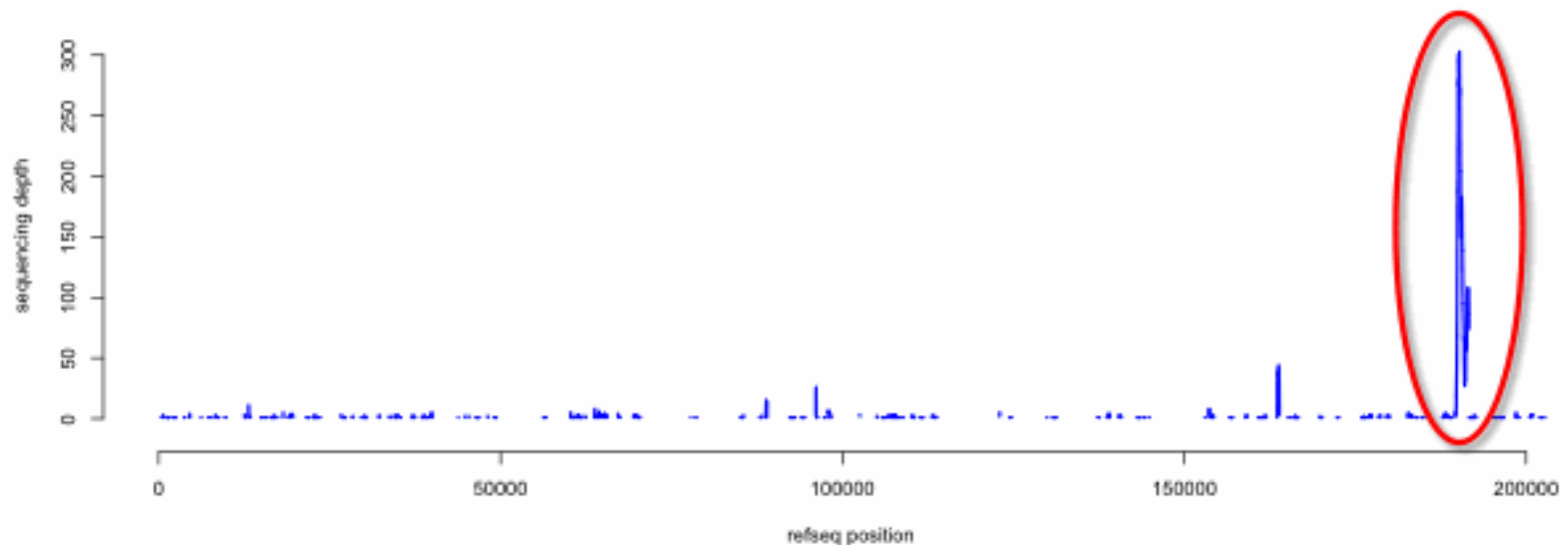
=> je grösser die Heterogenität, desto grösser der Sequenzieraufwand



Wieso grosse *Read*-Längen?

Probe UNC: unspezifischer *Mapping*-Hintergrund

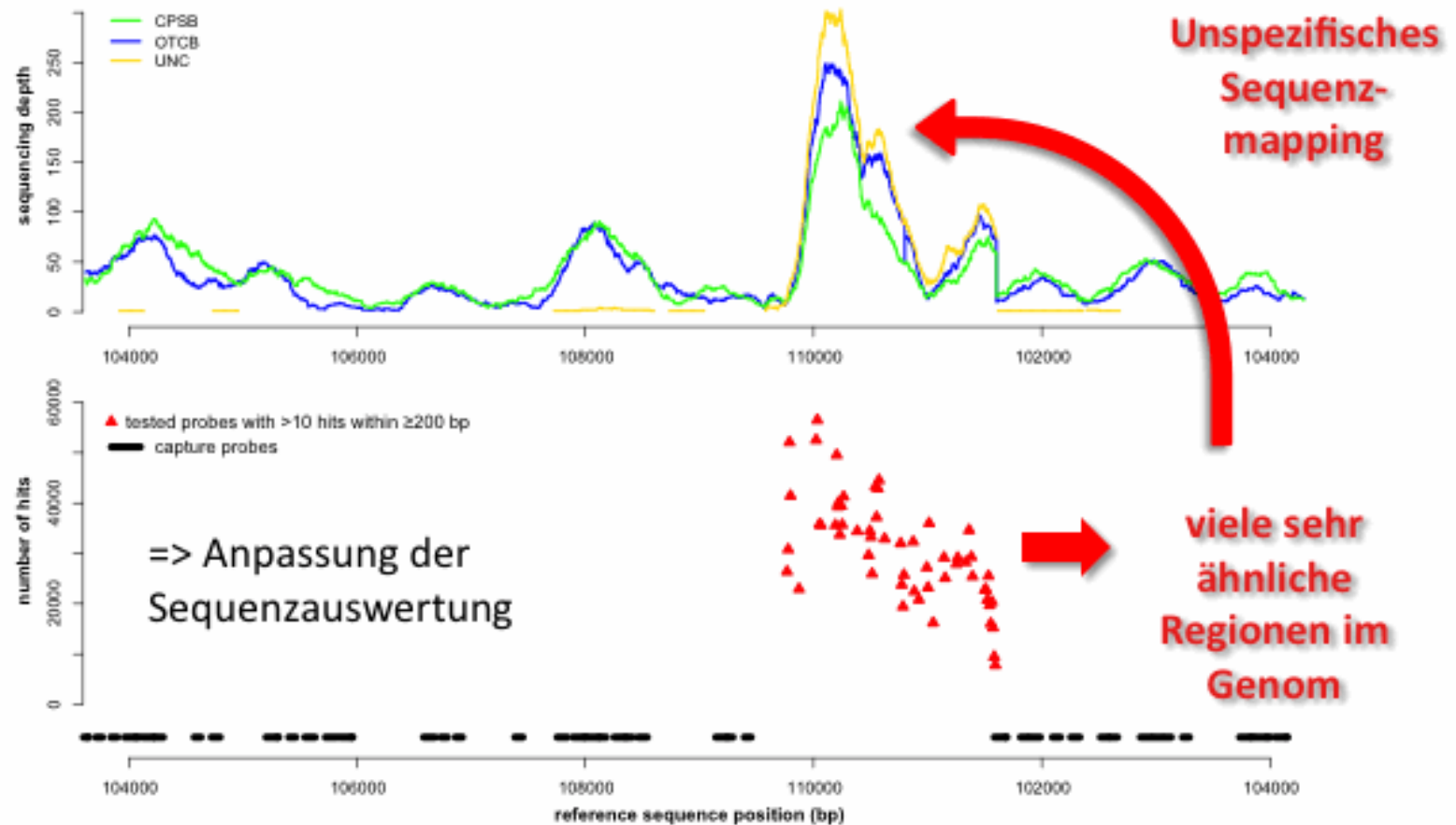
- **316677 bp** *mapped* innerhalb der Zielregion
- Mehr als durch Zufall erwartet werden!
- **78%** der Sequenzen *mapped* innerhalb einer 1800 bp -Region



Wieso grosse *Read*-Längen?

Probe UNC: unspezifischer *Mapping*-Hintergrund

- **316677 bp** *mapped* innerhalb der Zielregion
- Mehr als durch Zufall erwartet werden!
- **78%** der Sequenzen *mapped* innerhalb einer 1800 bp -Region



Wieso grosse *Read*-Längen?

Experiment 1

- Durchschnittliche *Read*-Länge: **311 bp**
- **89%** der *Reads mapped* einmalig im humanen Genom

Experiment 2

- Durchschnittliche *Read*-Länge: **219 bp**
- **58%** der *Reads mapped* einmalig im humanen Genom

→ Je länger die *Reads*, desto eindeutiger die Sequenzidentifikation, desto weniger unspezifisches Mapping (weniger FP), höhere *Coverage* und kleinere repetitive Abschnitte können sequenziert werden



Referenzdaten – Sensitivität / Spezifität

→ Referenzdaten ~35000 bp resequenziert mit **Sanger** Sequenzierung

Variant	deviant allele (%)	TP	FP	FN	TN
Depth ≥10					
Homozygous	≥80	8	0	0	19654
Heterozygous, cutoff A	20-80	21	4	0	19638
Heterozygous, cutoff B	30-70	20	1	1	19641
Depth ≥5					
Homozygous	≥80	16	0	0	34462
Heterozygous, cutoff A	20-80	23	8	2	34446
Heterozygous, cutoff B	30-70	22	4	3	34450

→ **Alle** homozygoten Varianten auch bei tiefer *Sequencing Depth* korrekt identifiziert



Referenzdaten – Sensitivität / Spezifität

→ Referenzdaten ~35000 bp resequenziert mit **Sanger**-Sequenzierung

Variant	deviant allele (%)	TP	FP	FN	TN
Depth ≥10					
Homozygous	≥80	8	0	0	19654
Heterozygous, cutoff A	20-80	21	4	0	19638
Heterozygous, cutoff B	30-70	20	1	1	19641
Depth ≥5					
Homozygous	≥80	16	0	0	34462
Heterozygous, cutoff A	20-80	23	8	2	34446
Heterozygous, cutoff B	30-70	22	4	3	34450

→ **Niedrige Sequencing Depth erhöht** Fehlerrate für heterozygote Varianten





Referenzdaten notwendig fürs „Tuning“ und zur Interpretation neuer Polymorphismen

Beobachtung 1 :

Hohe Fehlerrate bei Insertionen / Deletionen innerhalb oder nahe bei **homopolymeren Regionen** (≥ 4 mal die gleiche Base):

Nur 2 von 10 untersuchten Fälle konnten bestätigt werden!

Beobachtung 2 :

57 (92%) von 62 „neuen SNPs“, welche nicht mit Sanger-Sequenzierung untersucht wurden hatten einen Datenbankeintrag => Einengung der Kandidaten

Beobachtung 3:

Der einzige FP stammt wahrscheinlich von einem einzigen mutierten Amplikon ab, da alle *Reads* die gleiche Startposition im Sequenz-Alignment haben

=> Artefakt der Probenvorbereitung, kann bei der spezifischen Auswertung rausgefiltert werden.

Zusammenfassung

NGS ermöglicht relative grosse und komplexe DNA-Bereiche (mehrere Gene, viele Exons, nicht-kodierende Regionen etc.) für die Diagnostik zu erschliessen

Momentan relativ hohe Kosten (in diesem Bsp ca. 1500 CHF pro Patient) und lange Entwicklungszeiten (in diesem Bsp ca. 1 Jahr)

Sich rasch entwickelnde Technologie mit sinkenden Kosten und mit grossem Optimierungspotential (für unser Bsp. scheint eine Senkung der Kosten auf 1/3 möglich)

Die grössten Optimierungspotentiale liegen in der Homogenisierung der Coverage und der *Read*-Länge; nicht in der Erhöhung Anreicherungseffizienz

Eine Grosse Hürde sind mangelnde Referenzdaten (Spezifität, Sensitivität, Interpretation von neuen Varianten und Detektion von FP durch unspezifisches Mapping, Pseudogene und methodischen Arteakten).



Danke!

Ursula Amstutz

Gisela Andrey

Institute of Clinical Chemistry

Molecular Diagnostics

Inselspital Bern University Hospital

Johannes Häberle (Kinderspital Zürich)

Rolf Jaggi (DKF)

Dominic Suciu (CustomArray)

Marcelo Caraballo (CombiMatrix)



SWISS NATIONAL SCIENCE FOUNDATION



UNIVERSITÄT
BERN

