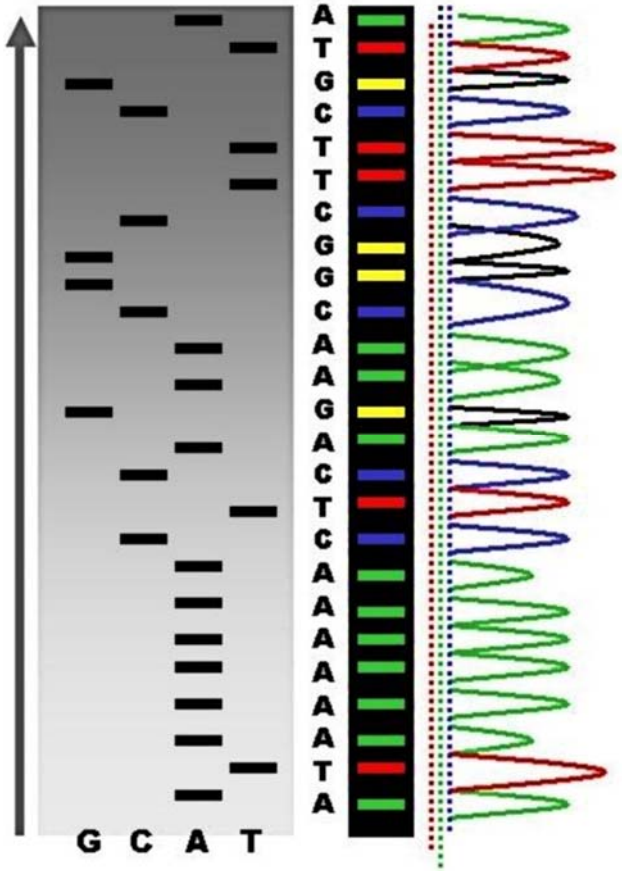
Archivierung von NGS-Daten – Artefaktsammlung oder Datenschatz?

Michael Nothnagel

Christian-Albrechts-Universität zu Kiel

Next-Generation Sequencing (NGS)



http://www.wikipedia.de/

- Nach Sanger Sequencing (1977) zweite Generation von Hochdurchsatz-Sequenzierungs-techniken
- Entwicklung seit den 1990er Jahren
- Seit 2005 verbreitete Anwendung, wiederholte Veränderungen und Verbesserungen
- Durchbruch in der Bestimmung genomischer Sequenzen

$\Rightarrow$ immer noch relativ neue Technik

NGS als Hoffnungsträger

- Direkte Untersuchung kausaler Varianten; mögliche Ablösung des Ansatzes indirekter Assoziationsstudien
- Teilweise Aufklärung der ‚missing heritability' in häufigen Erkrankungen durch Identifizierung seltener genetischer Varianten
- Gezieltere Tumortherapie mit Hilfe somatischer Mutationsprofile
- Umfassendere Analyse genetischer Information, u.a. durch Quantifizierung epigenetischer Modifikationen (Epigenomics) und intermediärer Genprodukte (Transcriptomics)
- und mehr…

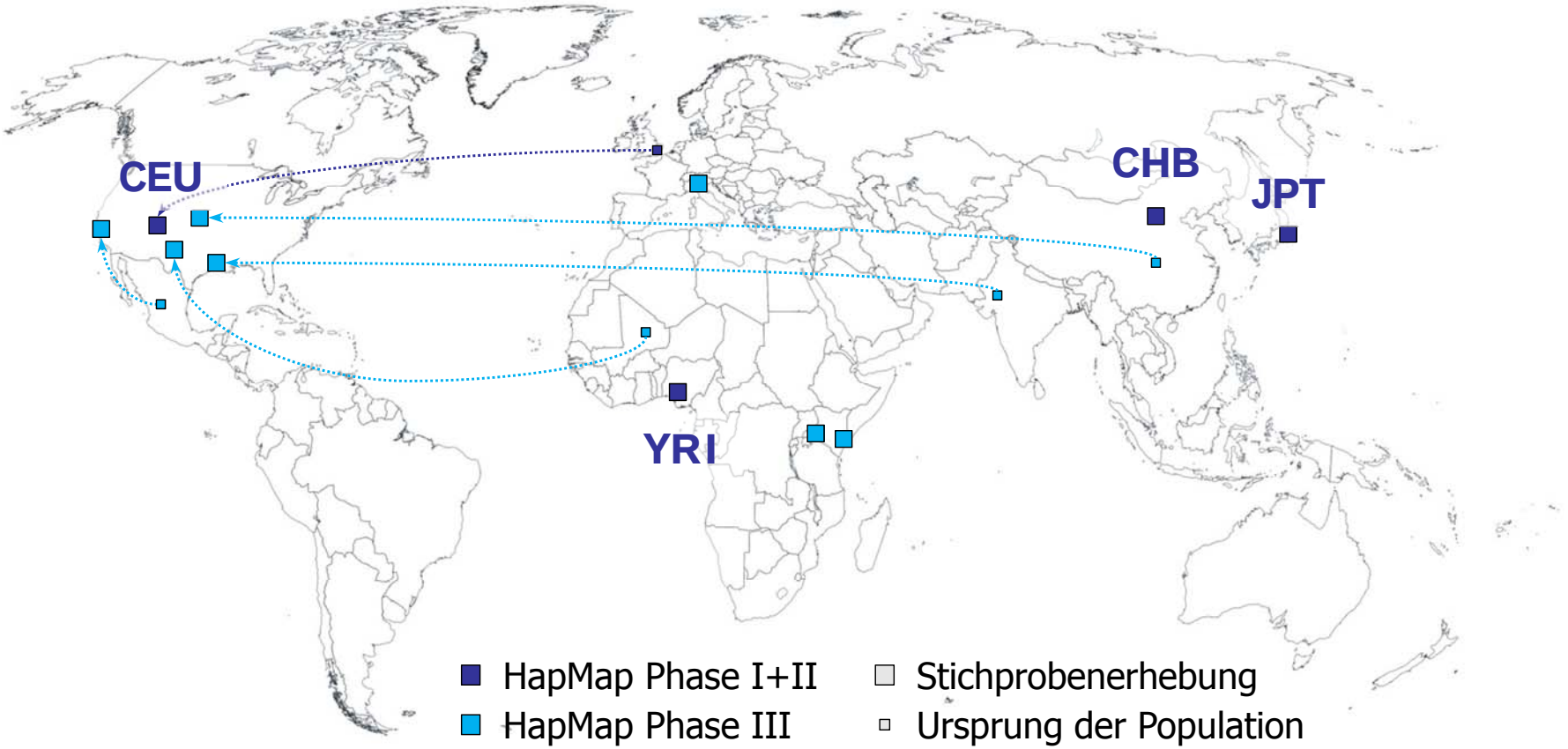
Neue Daten – Neue Fehler

- NGS-Daten sind fehlerbehaftet
- Natur der Fehler
  - im Vornherein häufig unbekannt (‚learning by doing', Erfahrungssammlung)
- Fehlerquellen für NGS-Daten (Auswahl):
  - Probleme beim Alignment von Reads (kurze Sequenzen teilweise unklaren genomischen Ursprungs)
  - einige Regionen sind nicht erreichbar (z.B. Pseudoautosomale Region der X/Y-Chromosomen, Repeats, Duplications etc.)
  - unterschiedlich hohe Abdeckung genomischer Regionen

Das HapMap-Referenz-Projekt

Ziel: Katalog häufiger genetischer Varianten in humanen Populationen

(basierend auf Genotypisierung, Frequenz≥5%; 3.1 Mill. SNPs in Phase I+II)

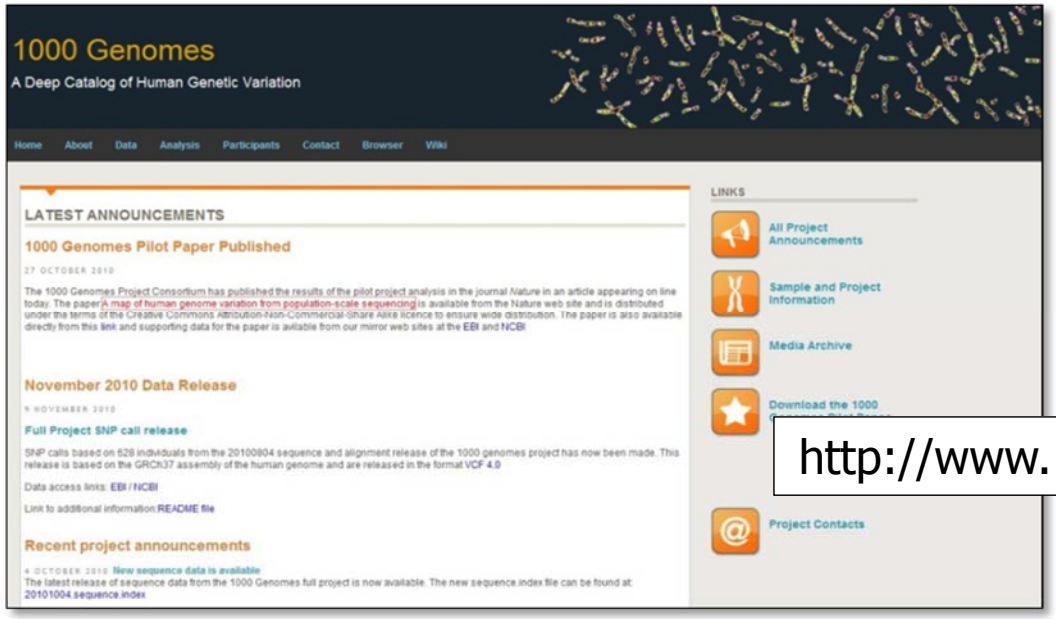


Referenz für: tagSNP-Auswahl, Genotyp-Imputation, Qualitätskontrolle, etc.

The International HapMap Consortium (2003,2005,2007) Nature

Das 1000-Genome-Projekt

Ziel: Katalog der meisten genetischen Varianten mit einer Frequenz ≥1% in den untersuchten Populationen (basierend auf Next-Gen-Sequenzierung)



http://www.1000genomes.org/

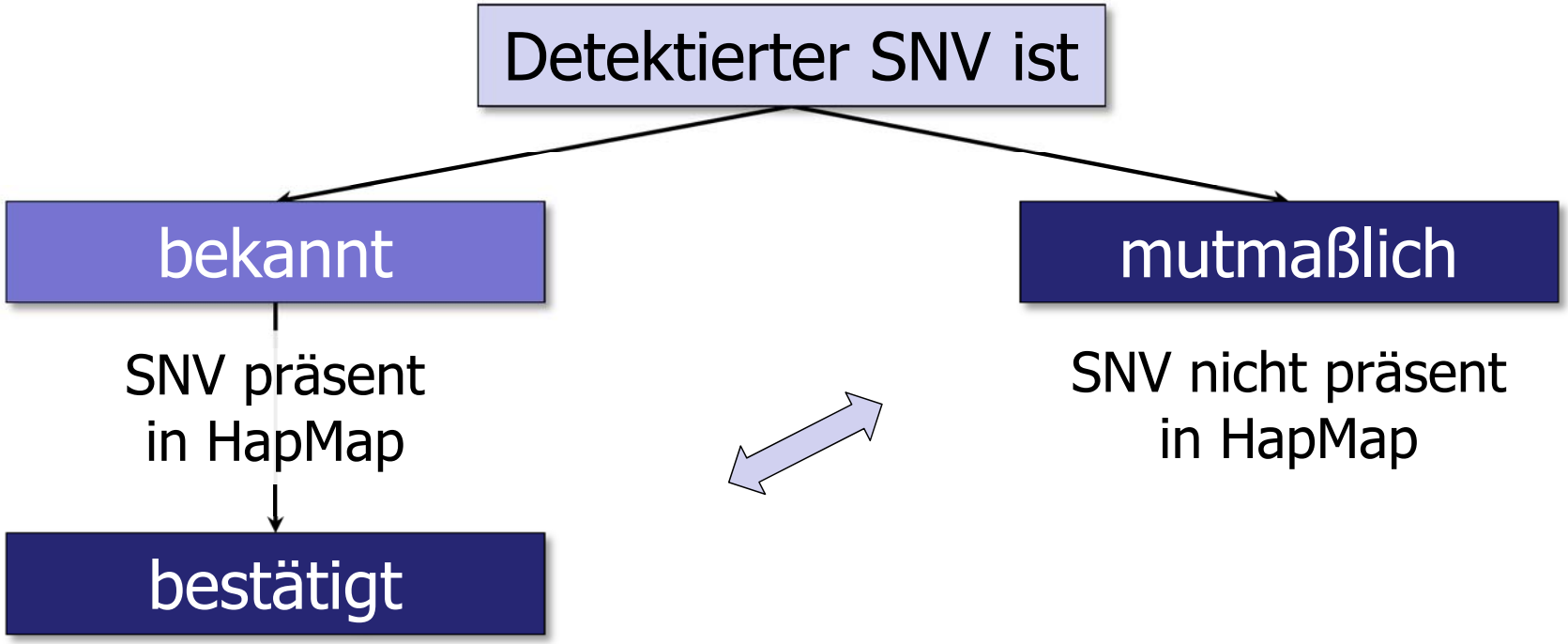
<u>Pilot 2 Projekt:</u>

Gesamtes Genom in 2 Trio-Familien (CEU, YRI) @ 20-60x Abdeckung

The 1000 Genomes Project Consortium (2010) Nature

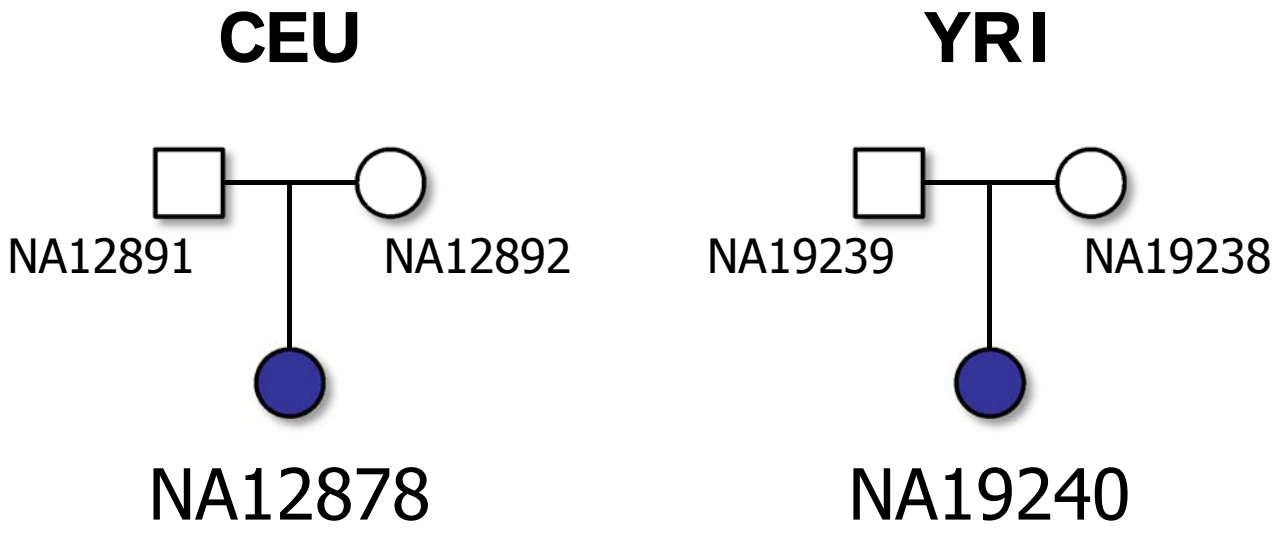
Fehler in NGS-basierter SNV-Detektion

Frage: Wie hoch ist der Anteil Falsch-Positiver unter neu detektierten (heterozygoten) SNVs?



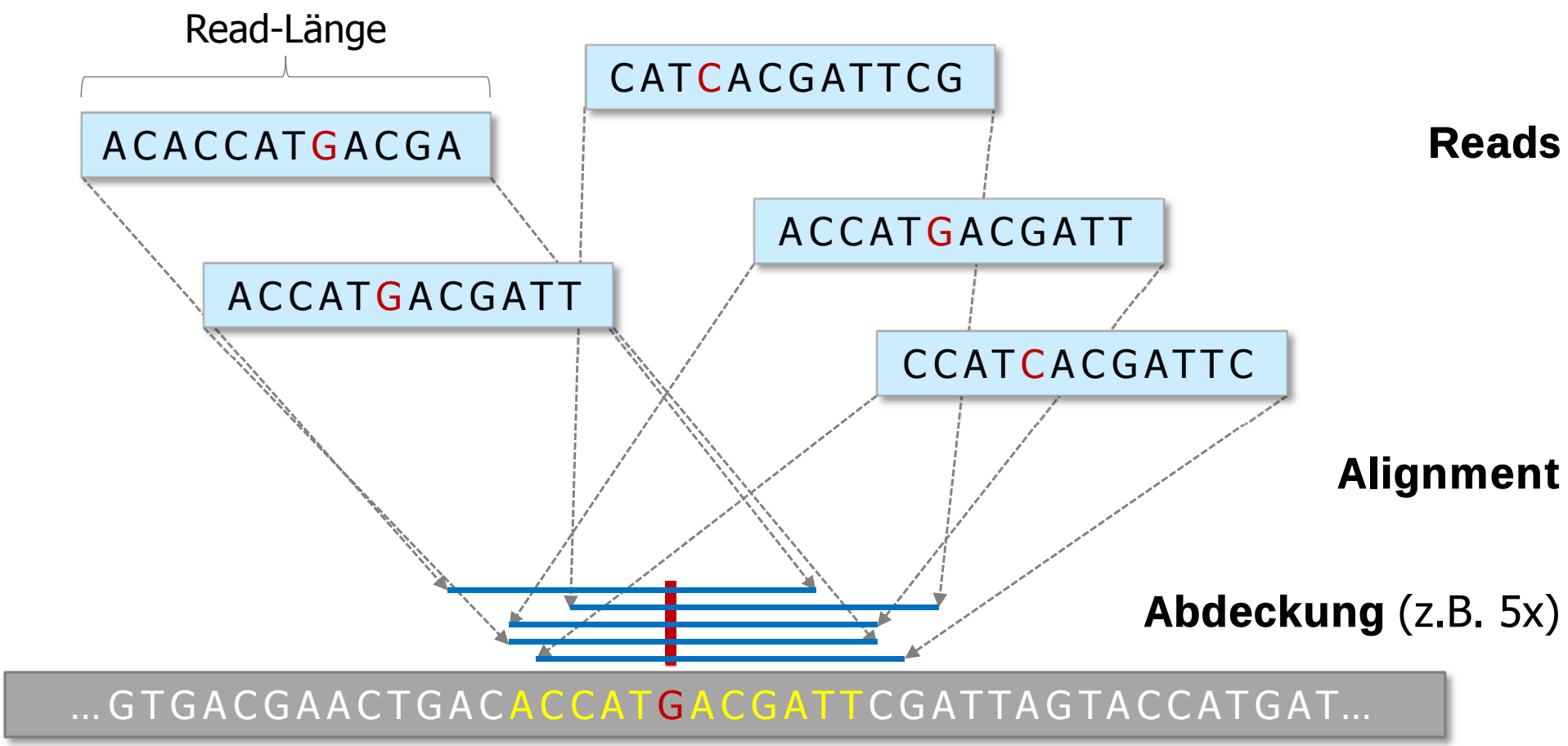
SNV: single-nucleotide variant

Analysierte Proben und Sequenzen



- Proben NA12878 and NA19240 wurden sequenziert mit je drei Technologien: 454 FLX™, GA IIx™ and SOLiD™
  [zu verschiedenen Zeiten, verschiedene Algorithmen und QC-Ansätze]
- Download der Daten alignter Sequence-Reads von der 1000 Genome Projekt-Webseite (Pilot 2 data set, Mai 2010)

NGS: SNV-Inferenz



Reads

Alignment

Abdeckung (z.B. 5x)

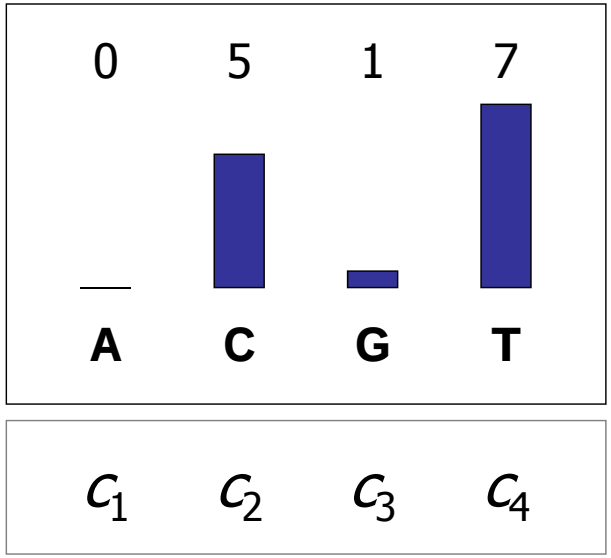
Referenzsequenz

G / C **Single-nucleotide variant (SNV)**

Software: SAMtools, GATK, diBayes, CASAVA, etc.

Verteilung allelspezifischer Reads

Allelspezifische Reads aus dem NGS



13

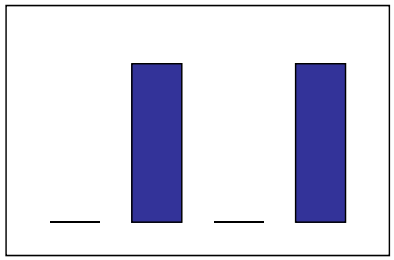
Abdeckung
[Coverage]

Read-Anzahl

Entropie: $H = -\sum p_i \log_2(p_i)$ $p_i = c_i / \Sigma_i c_i$

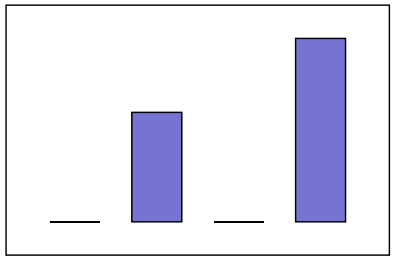
Entropie der allelspezifischen Reads

Anzahl allelischer Reads



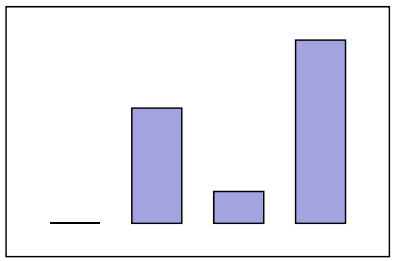
Gleichhäufige Reads für zwei Allele:

$H = 1$



Ungleichhäufige Reads für zwei Allele:

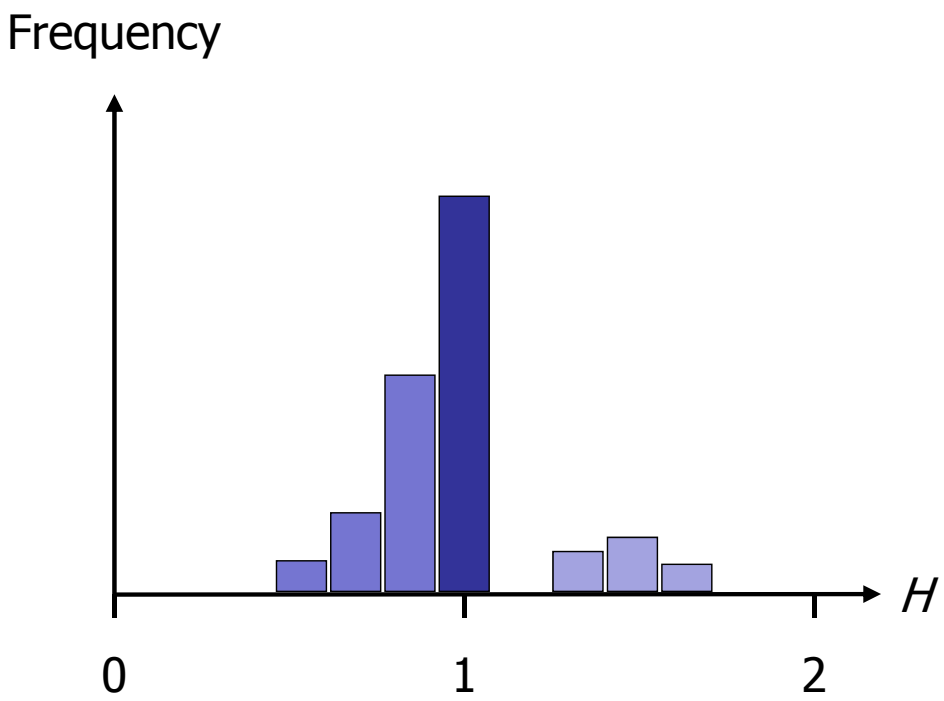
$H < 1$



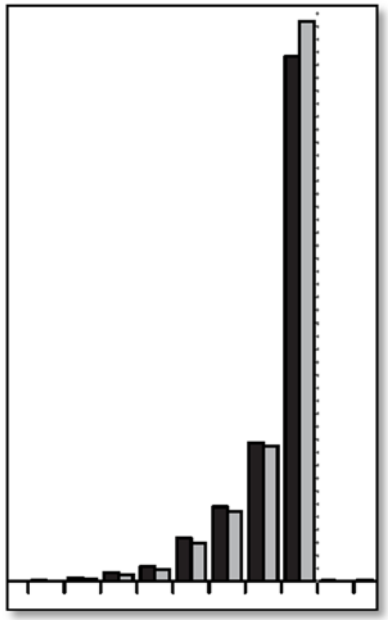
Präsenz von Reads für dritte / vierte Allele:

$H > 1$

Entropie-Histogramm (Beispiel)



Schätzung des Anteils falsch-positiver SNVs



■ Mutmaßliche SNVs

□ Bestätigte SNVs

Beobachtete Dichte von *H* für mutmaßliche SNVs

Mischung der Dichten heterozygoter und homozygoter (falsch-positiver) Genotypen

$$f_{\text{putative}} = (1-\alpha)\cdot f_{\text{het}} + \alpha \cdot f_{\text{hom}}$$

approximiert durch Gold-Standard!

normaler-weise nicht verfügbar

Schätzung für den Anteil falsch-positiver SNV-Detektionen:

$$\hat{\alpha} = \min\{\alpha : f_{\text{putative}}(x) \geq (1-\alpha)\cdot f_{\text{het}}(x), \forall x \in [0,1]\}$$

Bestätigte SNVs

Benutzung von Bins für die Schätzung

Übersicht über inferierte SNVs

1000-Genome-Projekt, Pilot 2, Chromosomen 1-22

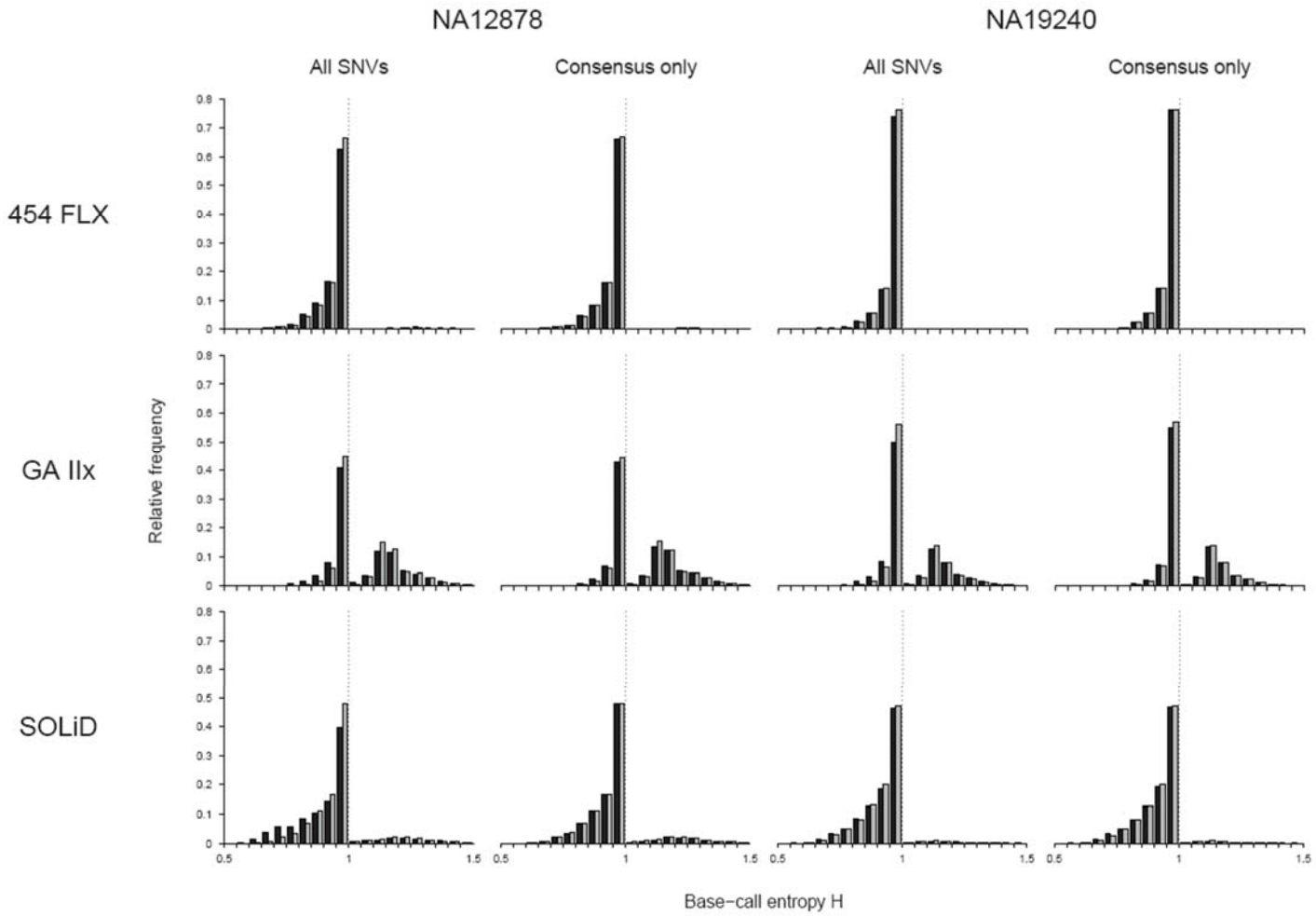
	NA12878 (CEU)		NA19240 (YRI)	
	Bekannte (Bestätigte) SNVs	Mutmaßliche SNVs	Bekannte (Bestätigte) SNVs	Mutmaßliche SNVs
454 FLX™	760,693 (724,548)	1,330,000	336,432 (319,106)	659,605
GA IIx™	821,017 (786,131)	1,126,727	892,372 (851,842)	1,816,994
SOLiD™	686,686 (651,873)	1,219,584	812,710 (777,840)	1,544,714
Konsens	609,429 (587,348)	631,533	300,237 (288,818)	420,570

- SNV-Inferenz: SAMtools (Li et al. 2009) mit Standard-Optionen
- SNV-Filter: quality score ≥20, read coverage ≤100
- Consensus: SNVs, die konkordant durch alle drei Plattformen gecallt wurden

Nothnagel et al. (2011) Hum Genet

Read-Entropie pro SNV

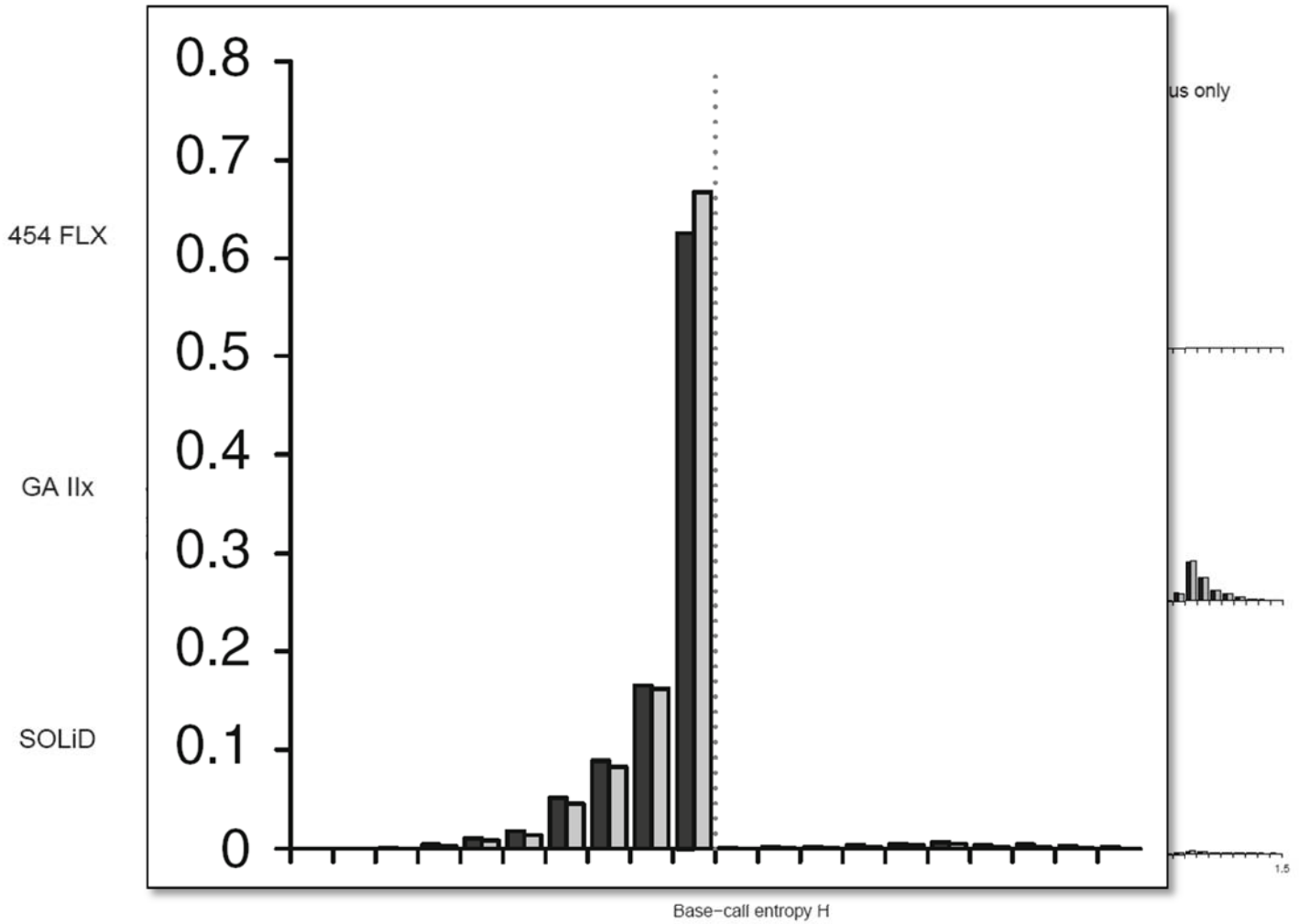
1000-Genome-Projekt, Pilot 2, Chromosomen 1-22



Nothnagel et al. (2011) Hum Genet

Read-Entropie pro SNV

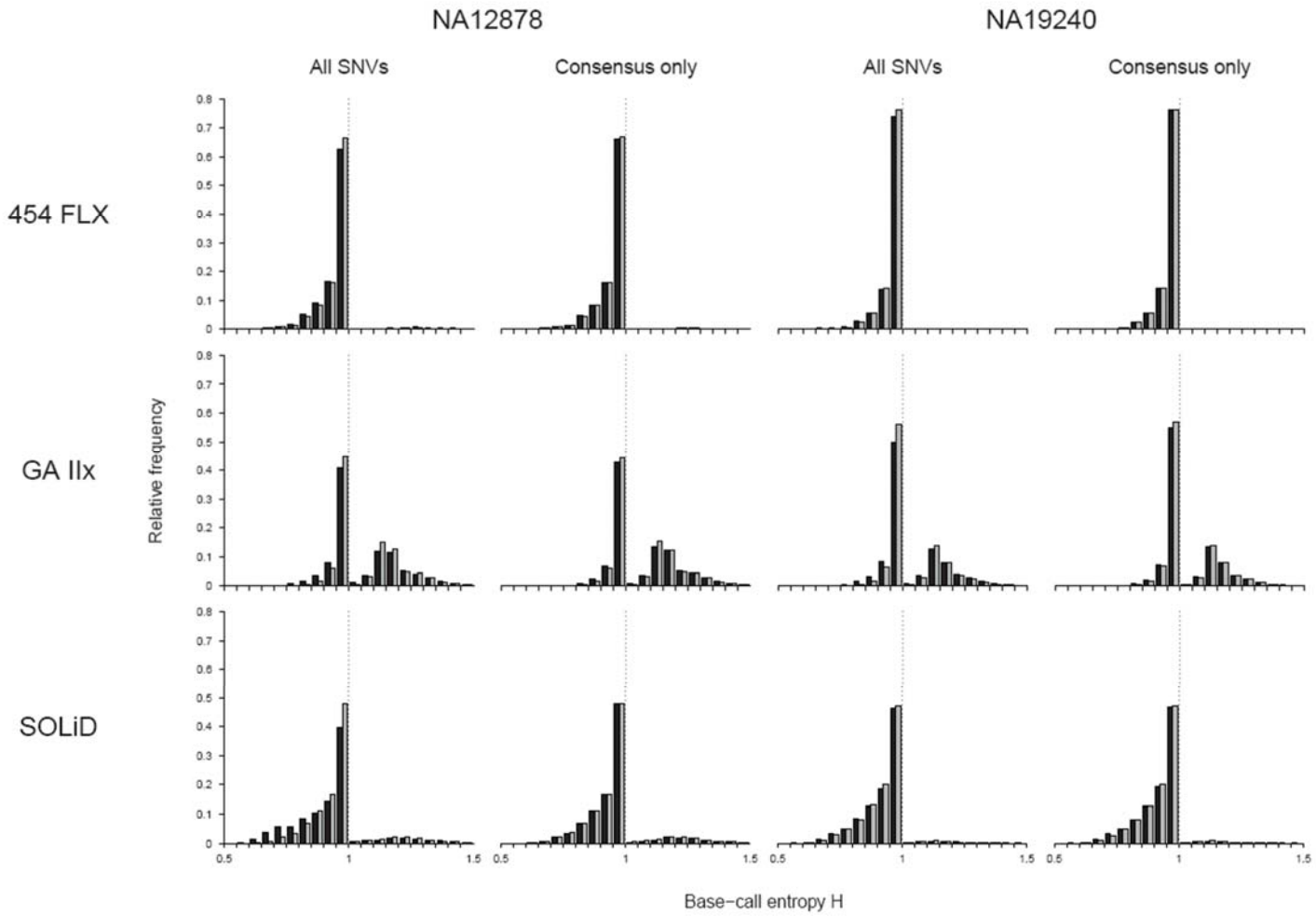
1000-Genome-Projekt, Pilot 2, Chromosomen 1-22



Nothnagel et al. (2011) Hum Genet

Read-Entropie pro SNV

1000-Genome-Projekt, Pilot 2, Chromosomen 1-22



Nothnagel et al. (2011) Hum Genet

Schätzungen des Anteils falsch inferierter SNVs

1000-Genome-Projekt, Pilot 2, Chromosomen 1-22

[%]	NA12878 (CEU)			NA19240 (YRI)		
	Alle SNVs	Konsens	P	Alle SNVs	Konsens	P
454 FLX™	6.3 (6.1-6.5)	0.7 (0.5-3.6)	$<10^{-4}$	2.9 (2.7-3.2)	2.6 (1.2-4.7)	0.08
GA IIx™	8.4 (8.0-8.7)	3.5 (3.1-3.9)	$<10^{-4}$	11.1 (10.9-11.3)	3.9 (3.5-4.3)	$<10^{-4}$
SOLiD™	17.1 (16.9-17.4)	0.8 (0.1-2.6)	$<10^{-4}$	7.3 (6.8-7.8)	4.0 (3.1-4.8)	$<10^{-4}$

P-Werte aus einem Permutationstest.

Nothnagel et al. (2011) Hum Genet

Fragen

I. Hilft Qualitätskontrolle (QC)?
   - ➢ Minimaler Schwellenwert für die Abdeckung ✗
   - ➢ Minimaler Schwellenwert für den Quality-Score ✓

II. Sind HapMap-Varianten ‚einfacher' zu sequenzieren? ✗
   - ➢ Untersuchung möglicher Unterschiede in der flankierenden Sequenz zwischen bestätigten und mutmaßlichen SNVs

III. Sind die Ergebnisse spezifisch für den Datensatz? ✗
   - ➢ Analyse der July 2010 Release des 1000-Genome-Projekts

IV. Sind die Ergebnisse spezifisch für den SNV-Calling-Algorithmus? ✗
   - ➢ Analyse von SNVs, die mit einem alternativen Algorithmus gecalled wurden: GATK

Fazit

- Unterschiedliche Fehlerprofile der Plattformen [spezifische Fehler und Anfälligkeiten]
- Risiko der Verfälschung von Analysen
- Konsens-Calls können Anteil irrtümlicher SNV-Detektionen reduzieren
- Public July 2010 Release erscheint in Teilen von schlechterer Qualität als Pilot2 Release (unklare QC?)

- NGS ist eine sich noch entwickelnde Technologie
  - mit Fehlern (bekannter und unbekannter Form)
  - ohne validierten Konsens über die Qualitätskontrolle der Daten
  - ohne Konsens über die ‚richtige' Form der Datenanalyse
  - ‚work in progress'

Datensicherung? Sicher!

- Datenarchivierung bietet Möglichkeiten zur Re-Analyse
  - Post-hoc-Überprüfung nach Berichten über Fehlerquellen
  - Erneute Analyse unterschiedlich gereinigter und analysierter Daten zur Herstellung von Vergleichbarkeit
  - Erneute Analyse der ursprünglichen Daten (Read-Daten) mittels korrigierter, neuer und verbesserter Verfahren
  - Detektion von Fälschungen
- Archivierung bis zu einem Konsens über Qualitätskontrolle und Analyseform notwendig (bei Standard-Analysen)
- Archivierung daher in den meisten Fällen wünschenswert und notwendig (in den nächsten 5 Jahren)

Danksagungen

Hum Genet
DOI 10.1007/s00439-011-0971-3

ORIGINAL INVESTIGATION

Technology-specific error signatures in the 1000 Genomes Project data

Michael Nothnagel · Alexander Herrmann · Andreas Wolf · Stefan Schreiber · Matthias Platzer · Reiner Siebert · Michael Krawczak · Jochen Hampe

M. Nothnagel and A. Herrmann contributed equally to this work.

http://www.ncbi.nlm.nih.gov/pubmed/21344269

- Andreas Wolf, Michael Krawczak
  (Christian-Albrechts-Universität zu Kiel)
- Alexander Herrmann, Stefan Schreiber, Rainer Siebert , Jochen Hampe
  (Universitätsklinikum Schleswig-Holstein, Campus Kiel)
- Mathias Platzer
  (Leibniz-Institut für Altersforschung, Jena)