

Methoden der Hochdurchsatzsequenzierung

Vorlesung 03

Sommersemester 2021

Phred quality score

From Wikipedia, the free encyclopedia

A **Phred quality score** is a measure of the quality of the identification of the **nucleobases** generated by automated **DNA sequencing**.^{[1][2]} It was originally developed for **Phred base calling** to help in the automation of DNA sequencing in the **Human Genome Project**. Phred quality scores are assigned to each nucleotide base call in automated sequencer traces.^{[1][2]} The **FASTQ format** encodes phred scores as ASCII characters alongside the read sequences. Phred quality scores have become widely accepted to characterize the quality of DNA sequences, and can be used to compare the efficacy of different sequencing methods. Perhaps the most important use of Phred quality scores is the automatic determination of accurate, quality-based **consensus sequences**.

Contents [hide]
1 Definition
2 History
3 Methods
4 Applications
5 Compression
6 References
7 External links

Definition [edit]

Phred quality scores *Q* are defined as a property which is logarithmically related to the base-calling error probabilities *P*.^[2]

$$Q = -10 \log_{10} P$$

or

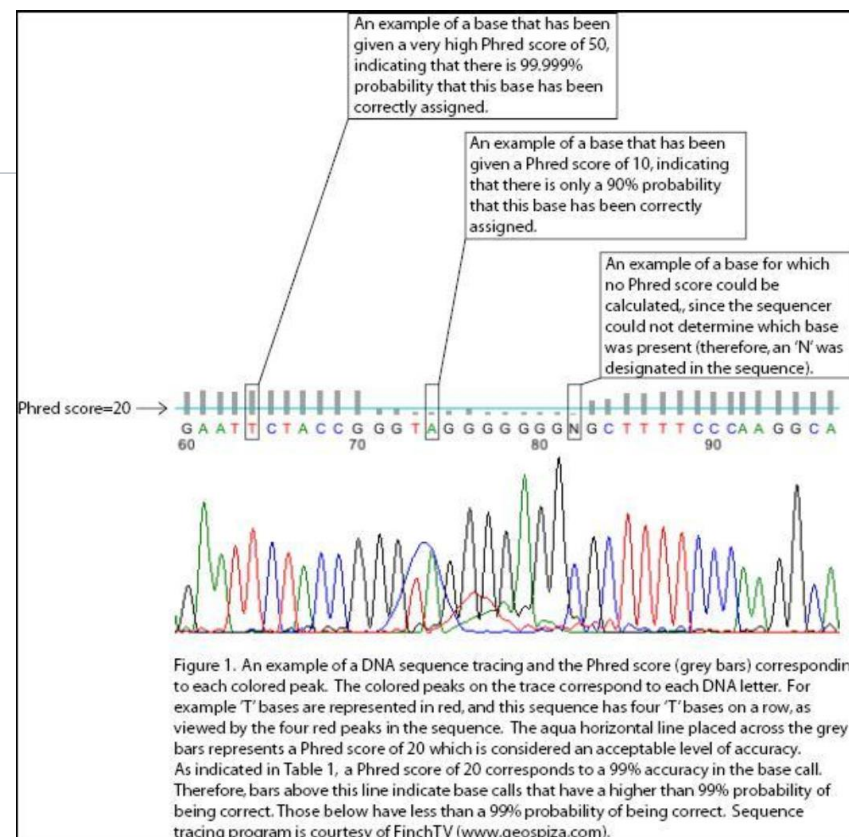
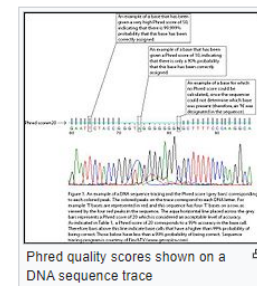
$$P = 10^{-\frac{Q}{10}}$$

For example, if Phred assigns a quality score of 30 to a base, the chances that this base is called incorrectly are 1 in 1000.

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

The phred quality score is the negative ratio of the error probability to the reference level of *P* = 1 expressed in **Decibel (dB)**.



Coverage (genetics)

From Wikipedia, the free encyclopedia

Coverage (or depth) in **DNA sequencing** is the number of unique reads that include a given **nucleotide** in the reconstructed sequence.^{[1][2]} **Deep sequencing** refers to the general concept of aiming for high number of unique reads of each region of a sequence.^[3]

 Contents [hide]
1 Rationale
1.1 Ultra-deep sequencing
1.2 Transcriptome sequencing
2 Calculation
3 Physical coverage
4 References

Rationale [edit]

Even though the sequencing accuracy for each individual nucleotide is very high, the very large number of nucleotides in the genome means that if an individual genome is only sequenced once, there will be a significant number of sequencing errors. Furthermore, many positions in a genome contain rare **single-nucleotide polymorphisms** (SNPs). Hence to distinguish between sequencing errors and true SNPs, it is necessary to increase the sequencing accuracy even further by sequencing individual genomes a large number of times.

Ultra-deep sequencing [edit]

The term "ultra-deep" can sometimes also refer to higher coverage (>100-fold), which allows for detection of sequence variants in mixed populations.^{[4][5][6]} In the extreme, error-corrected sequencing approaches such as Maximum-Depth Sequencing can make it so that coverage of a given region approaches the throughput of a sequencing machine, allowing coverages of >10⁸.^[7]

Transcriptome sequencing [edit]

Deep sequencing of **transcriptomes**, also known as **RNA-Seq**, provides both the sequence and frequency of RNA molecules that are present at any particular time in a specific cell type, tissue or organ.^[8] Counting the number of mRNAs that are encoded by individual genes provides an indicator of protein-coding potential, a major contributor to **phenotype**.^[9] Improving methods for RNA sequencing is an active area of research both in terms of experimental and computational methods.^[10]

Calculation [edit]

The average coverage for a **whole genome** can be calculated from the length of the original **genome** (*G*), the number of reads (*N*), and the average read length (*L*) as *N* × *L*/*G*. For example, a hypothetical genome with 2,000 base pairs reconstructed from 8 reads with an average length of 500 nucleotides will have 2× redundancy. This parameter also enables one to estimate other quantities, such as the percentage of the genome covered by reads (sometimes also called breadth of coverage). A high coverage in shotgun sequencing is desired because it can overcome errors in **base calling** and assembly. The subject of **DNA sequencing theory** addresses the relationships of such quantities.^[2]

Physical coverage [edit]

Sometimes a distinction is made between *sequence coverage* and *physical coverage*. Where sequence coverage is the average number of times a base is read, physical coverage is the average number of times a base is read or spanned by mate paired reads.^{[2][11][12]}

Read 1:	CGGATTACGTGGACCATG (read length of 18)
Read 2:	ATTACGTGGACCATGAATTGCTGACA
Read 3:	ACCATGAATTGCTGACATTCGTCA
Read 4:	TGAATTGCTGACATTCGTCA
Depth:	11122222222333344333333333332222221

An overlap of the product of three sequencing runs, with the read depth at each point indicated.

Read 1:	CGGATTACGTGGACCATG (read length of 18)
Read 2:	ATTACGTGGACCATGAATTGCTGACA
Read 3:	ACCATGAATTGCTGACATTCGTCA
Read 4:	TGAATTGCTGACATTCGTCA
Depth:	11122222222333344333333333332222221

[https://en.wikipedia.org/wiki/Coverage_\(genetics\)](https://en.wikipedia.org/wiki/Coverage_(genetics))

A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers

[Michael A Quail](#) , [Miriam Smith](#), [Paul Coupland](#), [Thomas D Otto](#), [Simon R Harris](#), [Thomas R Connor](#), [Anna Bertoni](#), [Harold P Swerdlow](#) & [Yong Gu](#)

BMC Genomics **13**, Article number: 341 (2012) | [Cite this article](#)

304k Accesses | **1162** Citations | **129** Altmetric | [Metrics](#)

Platform	Illumina MiSeq	Ion Torrent PGM	PacBio RS	Illumina GAIIx	Illumina HiSeq 2000
Instrument Cost*	\$128 K	\$80 K**	\$695 K	\$256 K	\$654 K
Sequence yield per run	1.5-2Gb	20-50 Mb on 314 chip, 100-200 Mb on 316 chip, 1Gb on 318 chip	100 Mb	30Gb	600Gb
Sequencing cost per Gb*	\$502	\$1000 (318 chip)	\$2000	\$148	\$41
Run Time	27 hours***	2 hours	2 hours	10 days	11 days
Reported Accuracy	Mostly > Q30	Mostly Q20	<Q10	Mostly > Q30	Mostly > Q30
Observed Raw Error Rate	0.80 %	1.71 %	12.86 %	0.76 %	0.26 %
Read length	up to 150 bases	~200 bases	Average 1500 bases**** (C1 chemistry)	up to 150 bases	up to 150 bases
Paired reads	Yes	Yes	No	Yes	Yes
Insert size	up to 700 bases	up to 250 bases	up to 10 kb	up to 700 bases	up to 700 bases
Typical DNA requirements	50-1000 ng	100-1000 ng	~1 µg	50-1000 ng	50-1000 ng

* All cost calculations are based on list price quotations obtained from the manufacturer and assume expected sequence yield stated.

** System price including PGM, server, OneTouch and OneTouch ES.

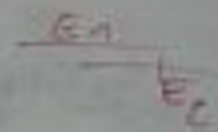
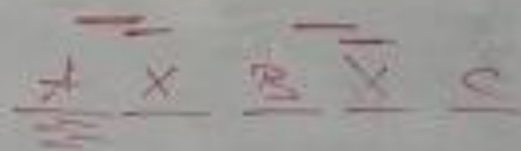
*** Includes two hours of cluster generation.

**** Mean mapped read length includes adapter and reverse strand sequences. Subread lengths, i.e. the individual stretches of sequence originating from the sequenced fragment, are significantly shorter.

Probleme

- * Satzwechselfehler
- * repeats
- * orientieren
- * Lücken (long+fatue)
- * Chimera
- * Zeit
- * Spracher

- Maximum likelihood
- Single / local alignment



Spindel → Gaps
Gaps

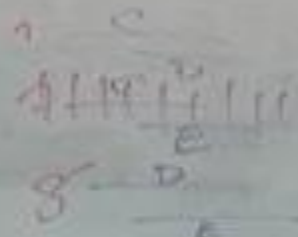
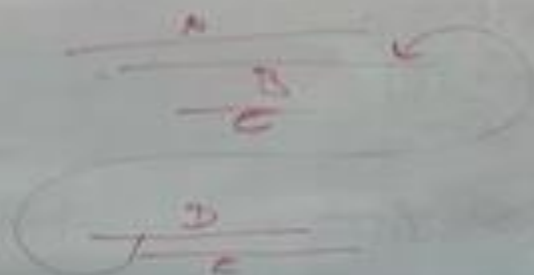
Chimera: pol
- subsequence
- long
- multiple

Wiederholung

GREEDY

1. perform alignment
2. max score
3. jump
4. repeat 1-3

Handwritten notes in a box:
- Chapter 10
- Greedy algorithm
- Dynamic programming



Maximum score
Single / two aligned seq.

