1. Qualitätskriterien Rohdaten

Tabelle 1 listet Qualitätskriterien für Rohdaten auf Basis der häufig genutzten Illumina und Oxford Nanopore (ONT) Technologien auf, die für eine Rekonstruktion des SARS-CoV-2 Genoms in Betracht gezogen werden können. Bei Verwendung alternativer Sequenziertechnologien sind vergleichbare Kriterien anzuwenden. Durch die Sequenzierung mindestens einer negativen Begleitprobe müssen Kontaminationen während der Prozessierung ausgeschlossen werden.

Tabelle 1. Qualitätskriterien für Rohdaten als Beispiel für Illumina und Oxford Nanopore (ONT) Technologien.

Qualitätsparameter	Illumina	ONT
Länge akzeptabler *Reads*[1]	≥ 30bp	≥ 200 bp
Durchschnittliche Qualität akzeptabler *Reads* (PHRED)	≥ 20	≥ 10

Nummer 1: Bei Verwendung amplicon-basierter Protokolle gilt zu beachten, die Länge akzeptabler Reads hinsichtlich der zu erwartenden Fragmentlängen anzupassen, um Primerdimere auszuschließen.

2. Qualitätskriterien rekonstruierte Genomsequenzen

Tabelle 2 listet zu erreichende Qualitätskriterien für erstellte SARS-CoV-2 Genomsequenzen unabhängig von der verwendeten Sequenziertechnologie auf. Die Qualität der zugrundeliegenden Rohdaten muss garantiert sein (siehe Punkt 1). Zudem muss beachtet werden, dass unerwünschte Sequenzen oder Sequenzbereiche wie Kontaminationen (z.B. humane Sequenzen), sequenzierte Adapter(teil)sequenzen und, im Falle amplicon-basierter Protokolle, Primersequenzen und -dimere vor dem *Variantcalling* ausgeschlossen werden. Die rekonstruierten Genomsequenzen müssen im IUPAC Nukleotidcode verfasst werden. Beispiele von Rekonstruktionspipelines, die die Einhaltung dieser Richtlinien ermöglichen, sind in Tabelle 3 genannt.

Tabelle 2. Qualitätskriterien für erstellte SARS-CoV-2 unabhängig von der verwendeten Sequenziertechnologie

Qualitätsparameter	Schwellenwert / Aktion
Identität zu NC_045512.2[2]	≥ 90%
Anteil N im rekonstruierten Genom	≤ 5%
Minimale lokale Sequenziertiefe ohne Filterung von PCR Duplikaten[3]	20
Minimale lokale Sequenziertiefe nach Filterung von PCR Duplikaten[3]	10
Informative Allelfrequenz[4]	≥ 90%
Frameshift-Mutationen[5]	besondere Absicherung

Nummer 2: Die relative Identität bezieht sich auf die Gesamtlänge der Referenzsequenz NC_045512.2. Als identisch werden ausschließlich übereinstimmende alignierte informative Positionen (A, T, G, C) gewertet.

Nummer 3: Positionen im rekonstruierten Genom, die von weniger als 20 *Reads* abgedeckt sind, müssen mit N maskiert werden. Sofern PCR Duplikate entfernt werden können, kann die minimale lokale Sequenziertiefe auf 10 reduziert werden.

Nummer 4: Informative Positionen (A, T, G, C) müssen durch 90% der alignierten *Reads* unterstützt bzw., im Falle von modellgestützter Basecall-Verfahren (z.B. bei Oxford Nanopore), mit vergleichbarer Sicherheit gewährleistet sein, ansonsten müssen weniger oder nicht-informative Platzhalter entsprechend des IUPAC Nukleotidcodes

Verwendung finden (z.B. R, Y , N). Im Falle gepaarter *Reads* muss eine strang-abhängige Verzerrung der Baseninformation (*Strandbias*) ausgeschlossen werden.

Nummer 5: Die Validität genomischer Variationen, die basierend auf der Annotation der Referenzsequenz NC_045512.2 zu Frameshifts innerhalb eines kodierenden Gens führen, muss am besten durch manuelle Inspektion des Alignments an den entsprechenden Positionen besonders abgesichert werden, um z.B. Lesefehler in homopolymeren Genomregionen auszuschließen.

Tabelle 3. Beispiele von Analysepipelines für die Generierung von SARS-CoV-2 Genomsequenzen aus Daten der häufig genutzter Illumina und Oxford Nanopore (ONT) Technologien.

Technologie	Git
Illumina	https://gitlab.com/RKIBioinformaticsPipelines/ncov_minipipe
ONT	https://github.com/replikation/poreCov