

Transcriptomics: single cell RNAseq

Bioinformatics & Omics

Salut!



- Doktoraarbeit an der UZH
- Postdocs at UZH und UniL
- Dozentin an Columbia University New York, USA
- seit 2018 an der ZHAW

Computational Genomics Lab:

Prof Dr. Maria Anisimova



Research Centre for Bioinformatics

We develop practical solutions at the interface of biology, medicine and computational sciences.

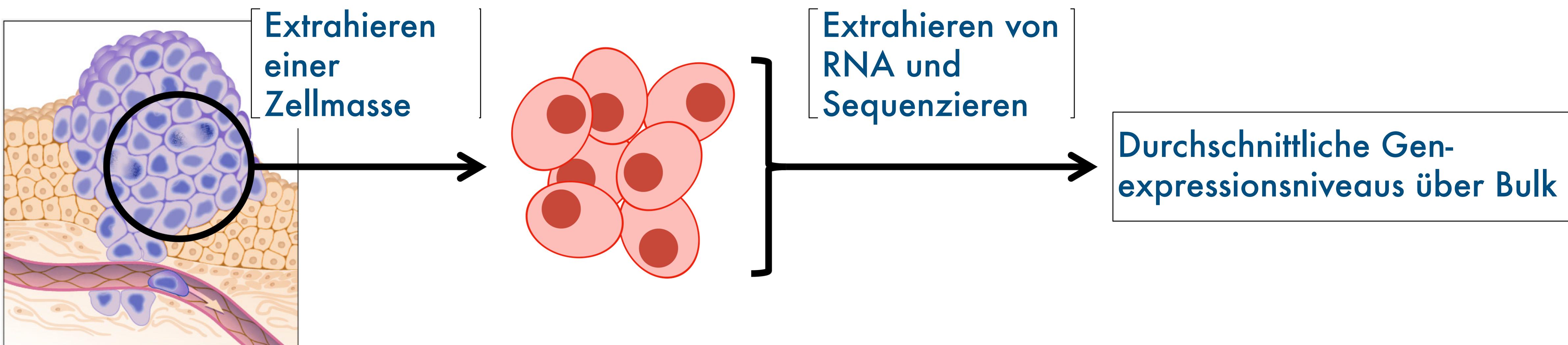
Bioinformatik-Fachwissen Frage

Heb deine Hand/halte sie hoch, wenn du

- schon einmal von Bioinformatik gehört hast,
- schon einmal von DNA-/RNA-Sequenzen gehört hast,
- schon einmal von einer Sequenzdatei (Fasta) gehört hast,
- eine Sequenzdatei geöffnet hast
- ein Online-Bioinformatik-Tool verwendet hast
- ein paar Zeilen programmiert hast
- eine Bioinformatik-Analyse durchgeführt hast
 - in R
 - in Python
- nervös bist
- aufgeregzt bist :)

Bulk-Transkriptomik

In den frühen Phasen der Next-Generation-Sequenzierung basierten alle Protokolle auf Bulk-Sequenzierungsansätzen. Bei der Bulk-Sequenzierung wird ein Teil des interessierenden Gewebes extrahiert, aus dem die DNA oder RNA isoliert und als einzelne Einheit analysiert wird. Das bedeutet, dass die Expressionsniveaus der Gene die durchschnittlichen Expressionswerte aller Zellen in der analysierten Probe widerspiegeln.

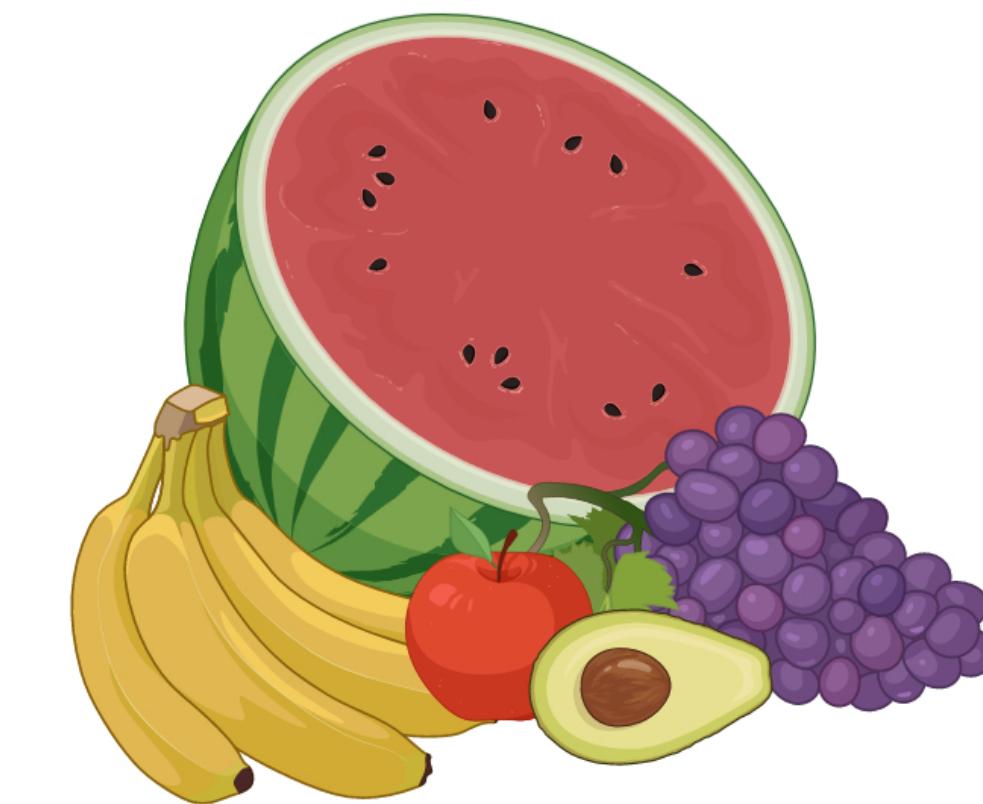


Bulk- vs. Einzelzell-Transkriptomik

- Bulk RNA-seq: Durchschnitt über viele Zellen
- Einzelzell-RNA-seq (scRNAseq): Jede Zelle einzeln analysiert



**Bulk
RNA-seq**



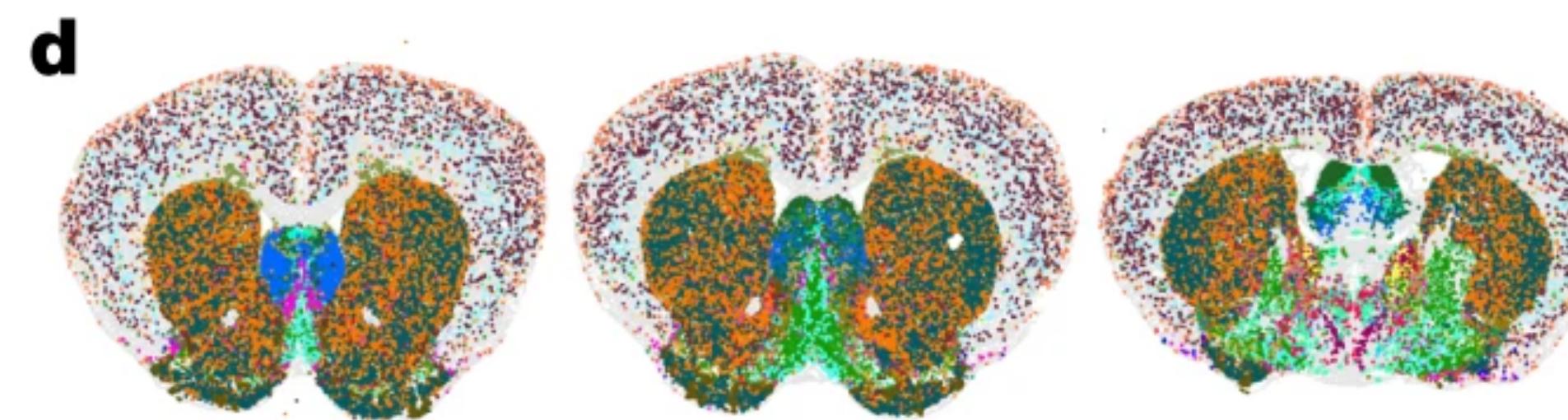
**Einzelzell
-RNA-seq**

Warum Einzelzell?

Bulk RNA-seq ist für transkriptomische Analysen nicht ideal, da in jeder Probe in der Regel verschiedene Zelltypen mit unterschiedlichen Genexpressionsprofilen vorhanden sind.

Verschiedene Zelltypen in einem Organ

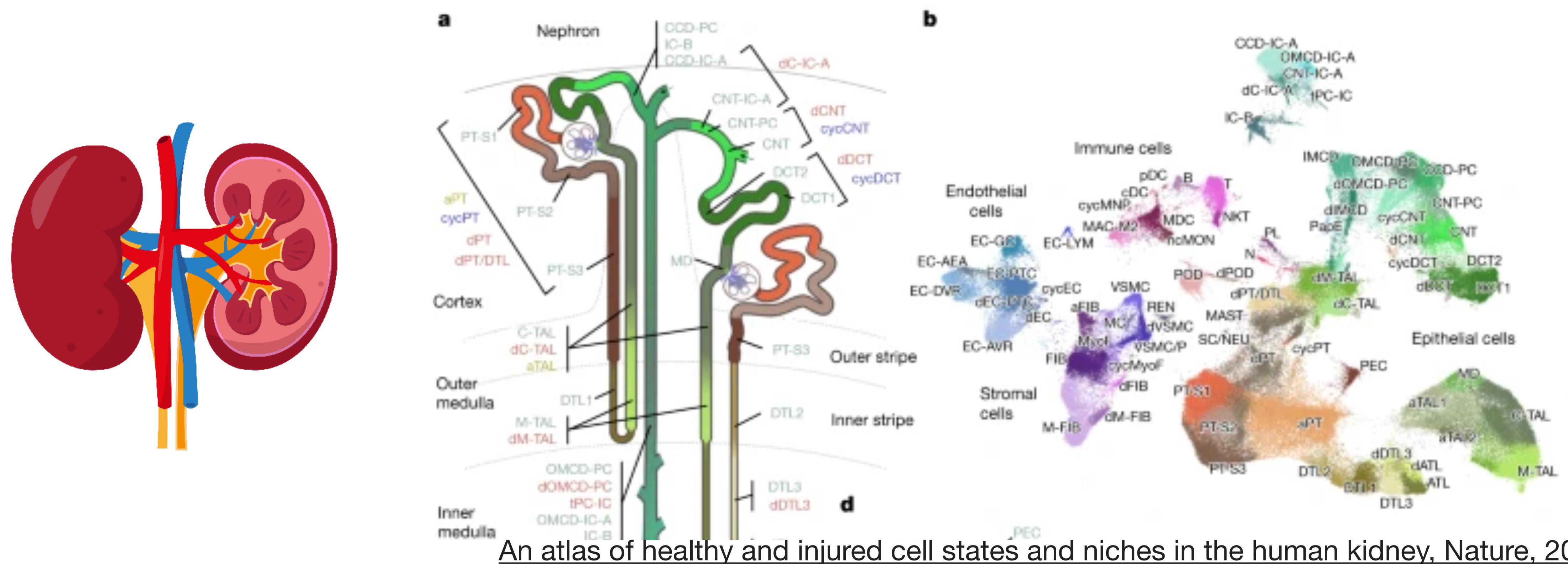
Nervenzellen und Nicht-Nervenzellen im Gehirn: zur Erforschung der Gehirnfunktion und neurologischer Störungen



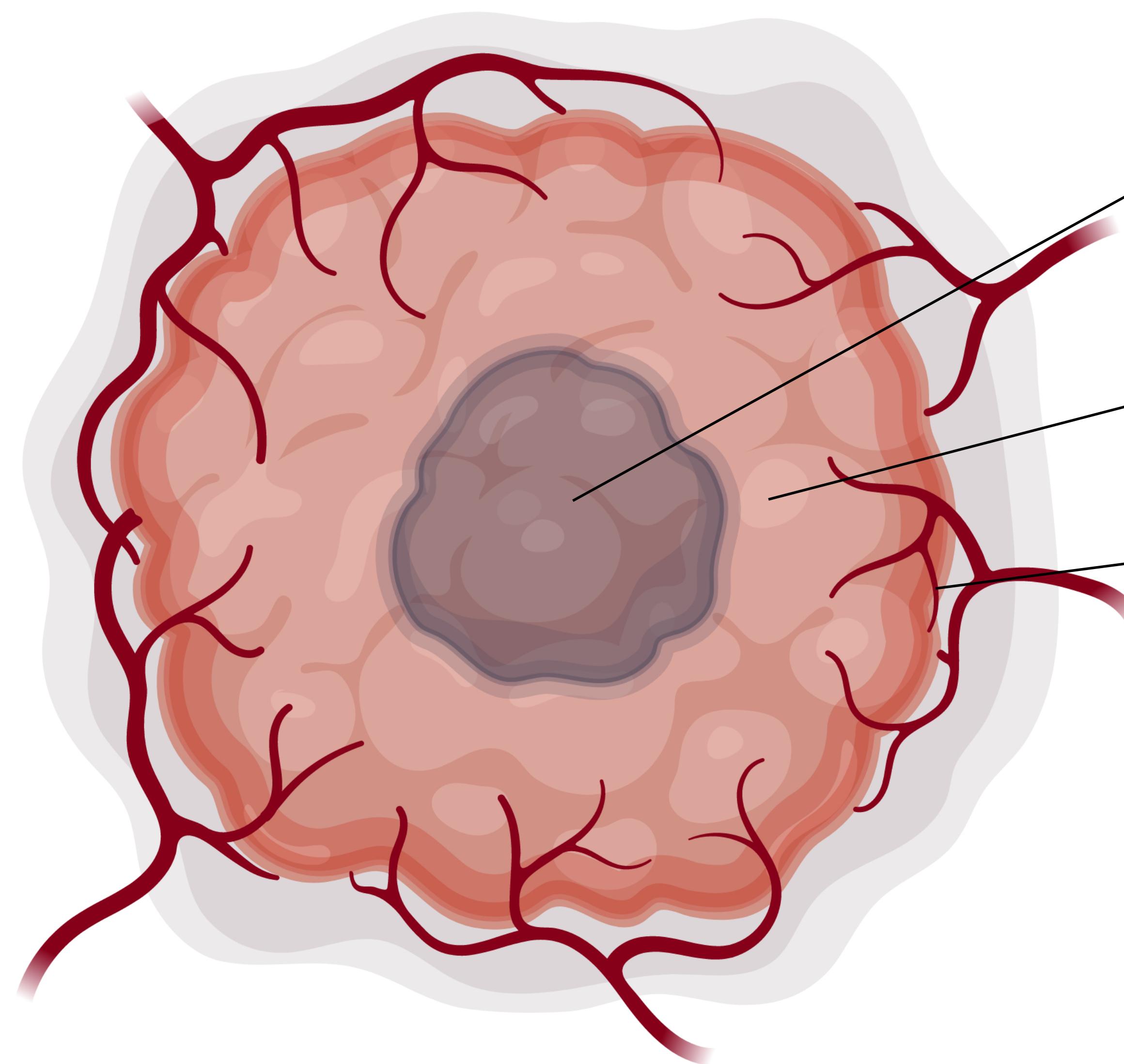
A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain, Nature, 2023

Verschiedene Zelltypen in einem Organ

Niere: Rinde, Mark usw. weisen alle unterschiedliche Zelltypen (Epithelzellen, Schichtzellen, Immunzellen usw.) und unterschiedliche Genexpressionen auf.



Krebsforschung



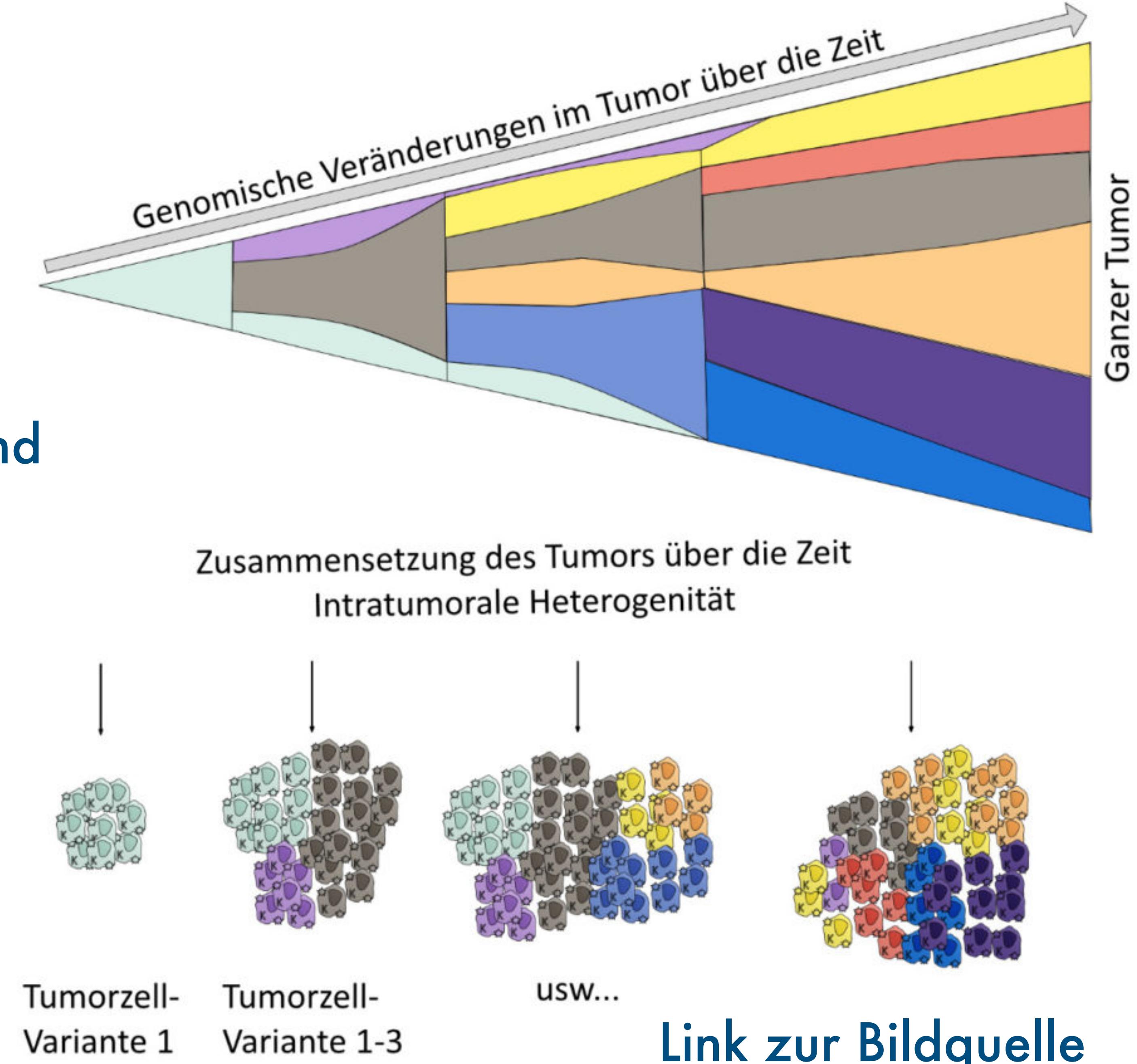
Nekrotischer Kern
Mangel an Sauerstoff und Nährstoff

Hypoxischer Bereich
Niedriger Sauerstoffgehalt und
geringe Nährstoffversorgung

In der Nähe der Blutgefäße
Hoher Sauerstoff- und Nährstoffgehalt

Krebsforschung

Tumorheterogenität:
Unterschiede zwischen
Krebszellen innerhalb eines
einzelnen Tumors. Entscheidend
für die Diagnose und die
Entwicklung und Bestimmung
des Ansprechens auf die
Therapie.



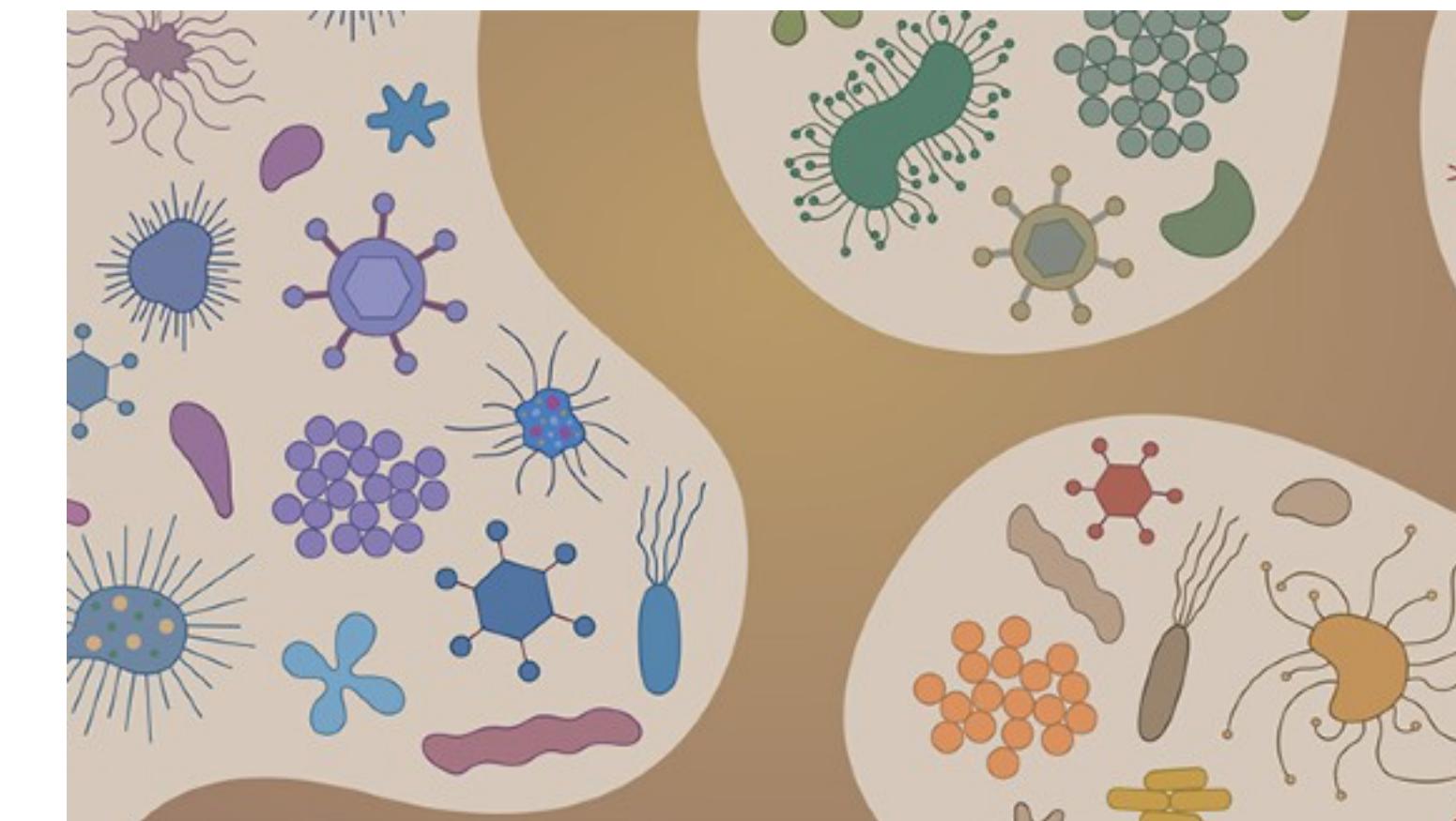
Forensik

- Identifizieren, aus welchem Gewebe die Hautzellen stammen (Hand, Geschlechtsorgane usw.)
- Identifizieren des Zelltyps (Blut, Speichel, Haut)
- Identifizieren, zu wem der Zelltyp gehört (Opfer, Täter, irrelevant)



Mikrobiom

Anwendungen in den Bereichen Lebensmittelsicherheit, Lebensmittelproduktion, schnelle Krankheitsdiagnostik, Identifizierung und Rückverfolgung von Krankheitserregern, Qualitätskontrolle, Entdeckung neuer Antibiotika sowie zur Untersuchung der Auswirkungen des Mikrobioms auf Gesundheit und Krankheit, beispielsweise bei entzündlichen Darmerkrankungen, Forensik



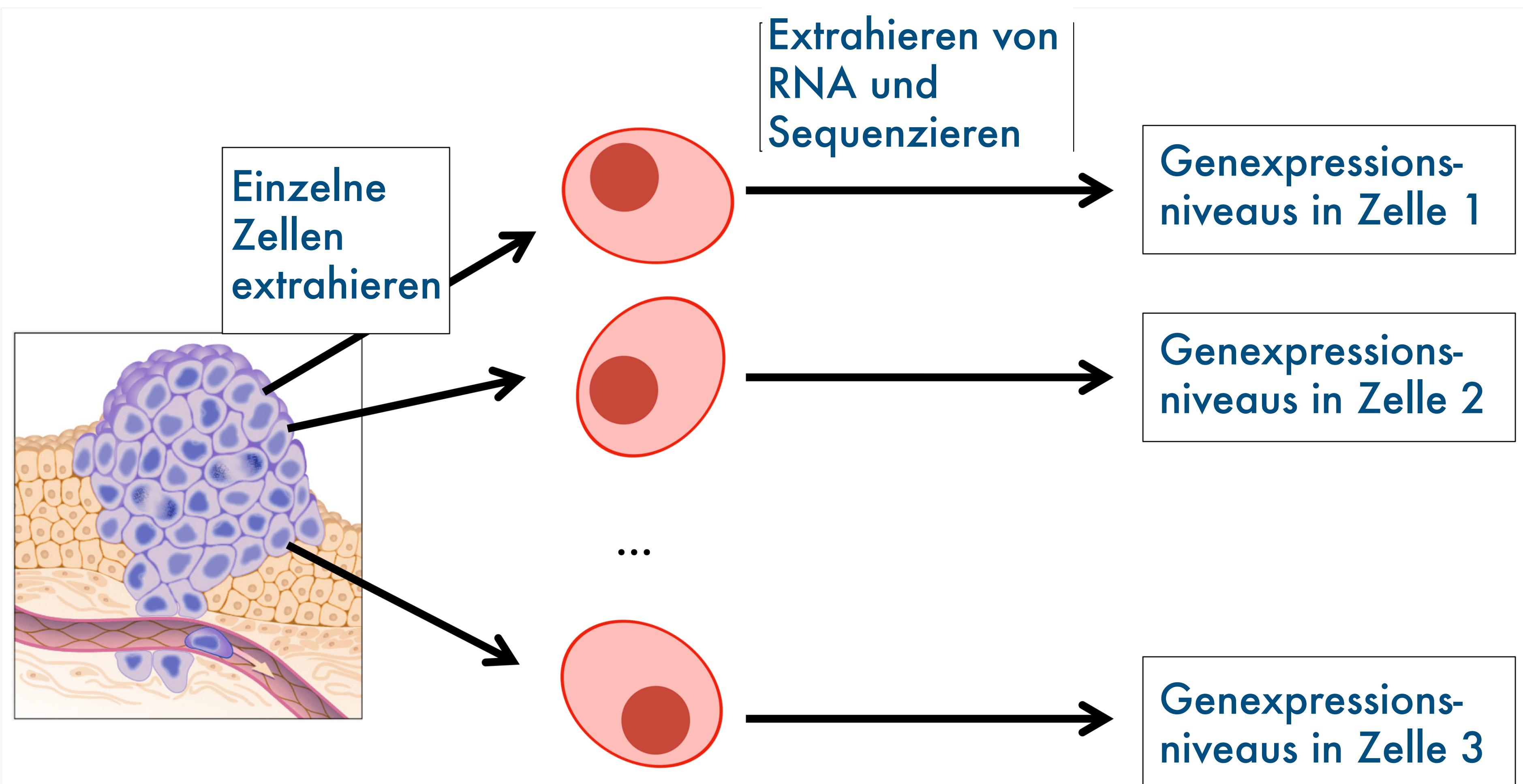
Bulk- vs. Einzelzell-Transkriptomik

Bulk RNA-seq ist für transkriptomische Analysen nicht ideal, da in jeder Probe in der Regel verschiedene Zelltypen mit unterschiedlichen Genexpressionsprofilen vorhanden sind.

Bei der Einzelzell-RNA-Sequenzierung (scRNAseq) werden einzelne Zellen aus einem Gewebe oder einer Probe von Interesse extrahiert. Aus jeder dieser Zellen wird die RNA extrahiert und separat sequenziert. Der resultierende Datensatz enthält nun Informationen zu einzelnen Zellen.

Einzelzell-Transkriptomik

Bei scRNAseq werden einzelne Zellen aus einer Probe von Interesse extrahiert. Aus jeder dieser Zellen wird die RNA extrahiert und separat sequenziert. Der resultierende Datensatz enthält nun Informationen zu einzelnen Zellen.



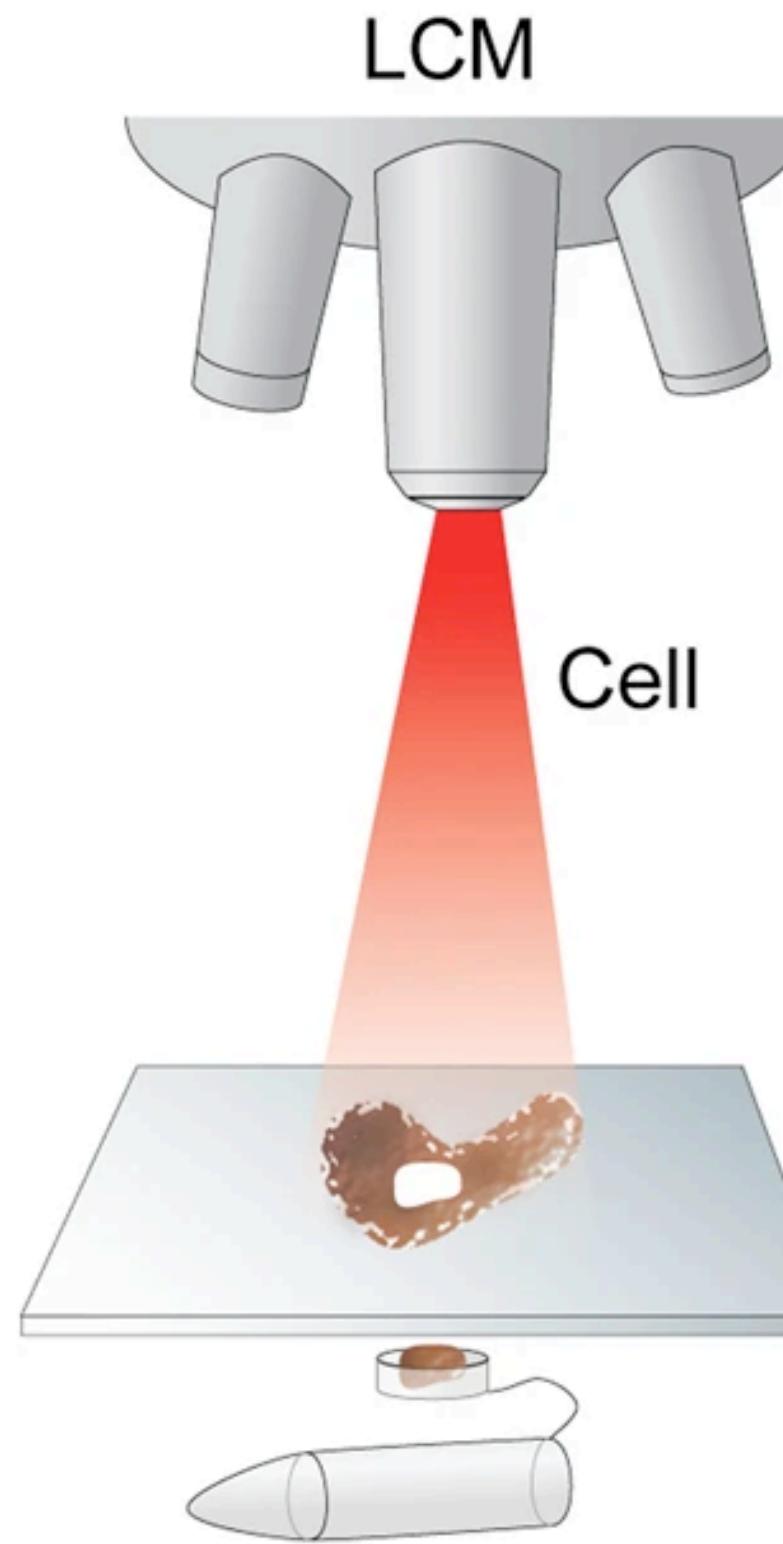
Herausforderungen bei der Einzelzell-Transkriptomik

Im Vergleich zu Bulk-RNA-Seq-Daten ist die Arbeit mit scRNA-Seq **besonders schwierig**.

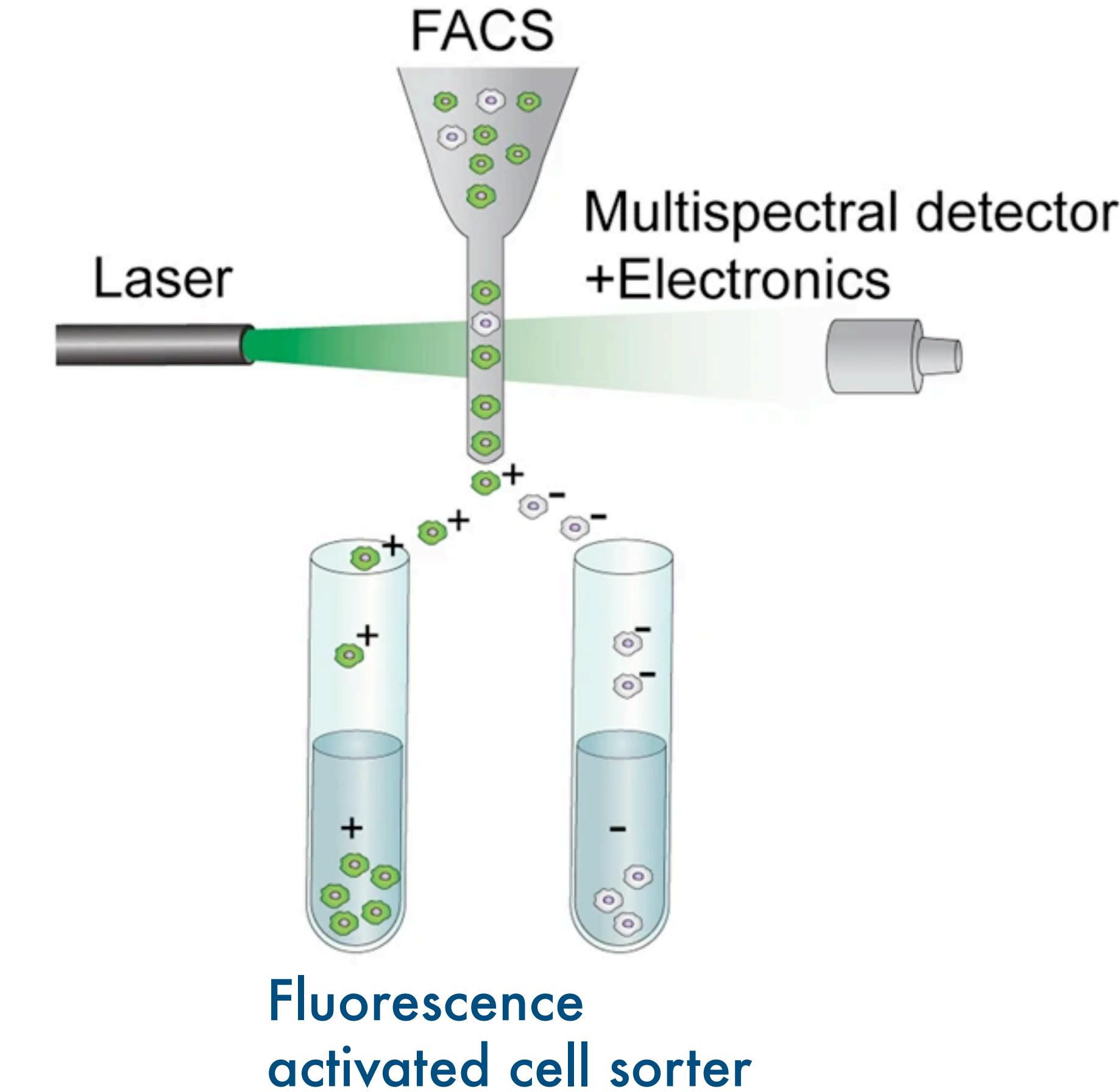
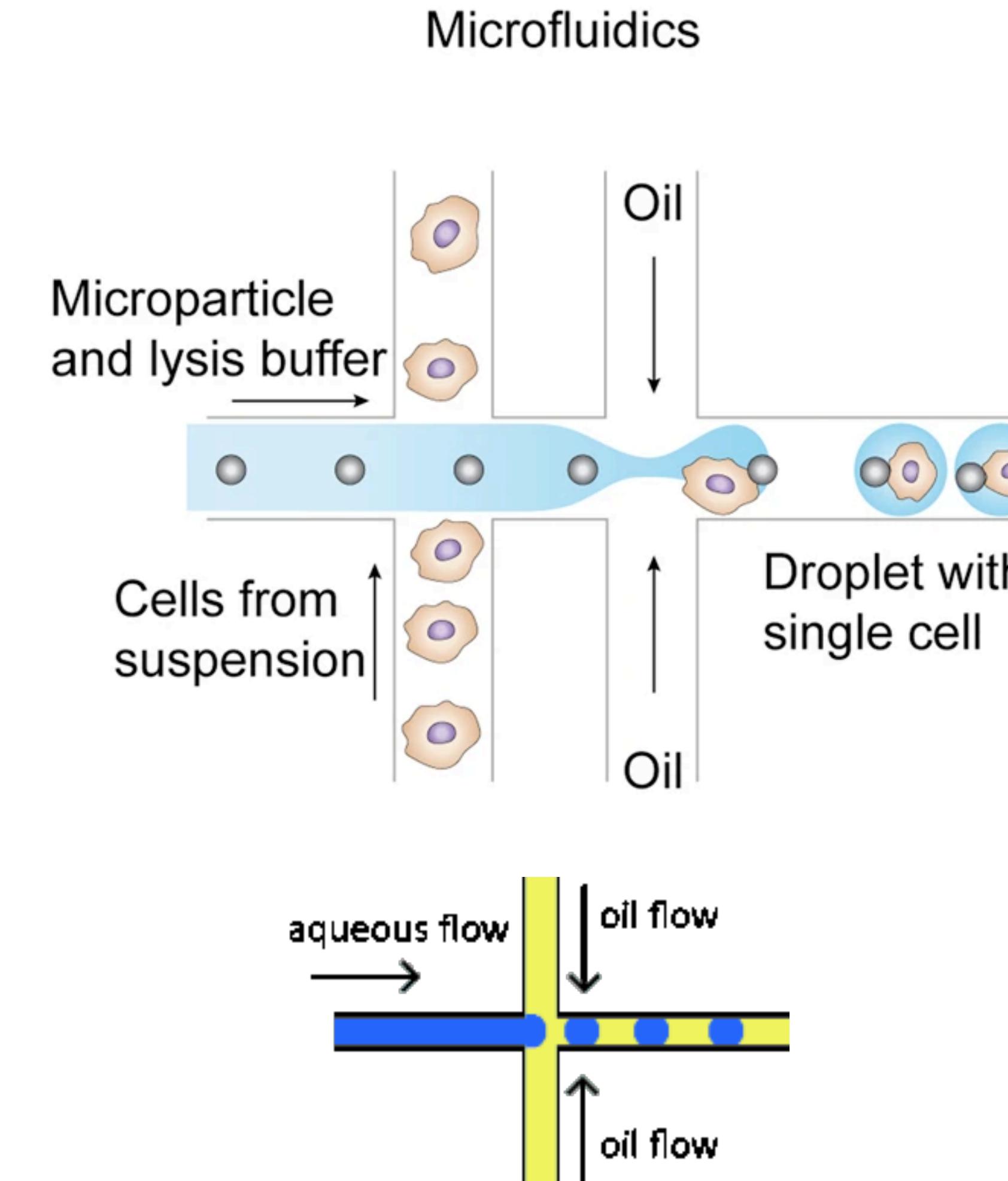
Bei scRNA-Seq sequenzieren Sie **sehr geringe Mengen an RNA-Ausgangsmaterial**, wodurch die Messungen anfälliger für **Artefakte** sind.

Darüber hinaus weisen einzelne Zellen eine natürliche **biologische Variabilität** in der Genexpression auf, was zu erheblichen Schwankungen in den aus Einzelzellanalysen abgeleiteten Schätzungen führt. Diese inhärente Heterogenität trägt zu einer erhöhten Unsicherheit bei der Beurteilung der Transkriptionszustände bei. Diese Unsicherheit kann jedoch durch die **Sequenzierung von mehr Zellen** gemindert werden, was ein umfassenderes Verständnis der gesamten Genexpressionslandschaft ermöglicht.

Technologien zur Erfassung einzelner Zellen



Laser Capture
Microdissection

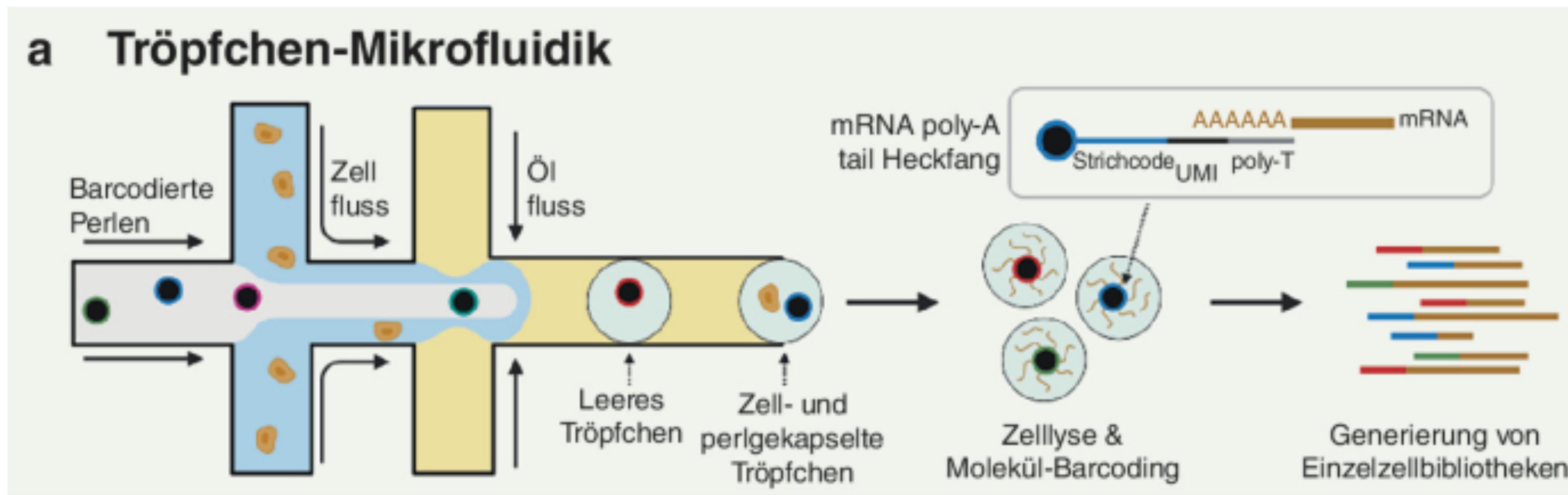


Fluorescence
activated cell sorter

Hwang, B., et al.. *Exp Mol Med* 2018

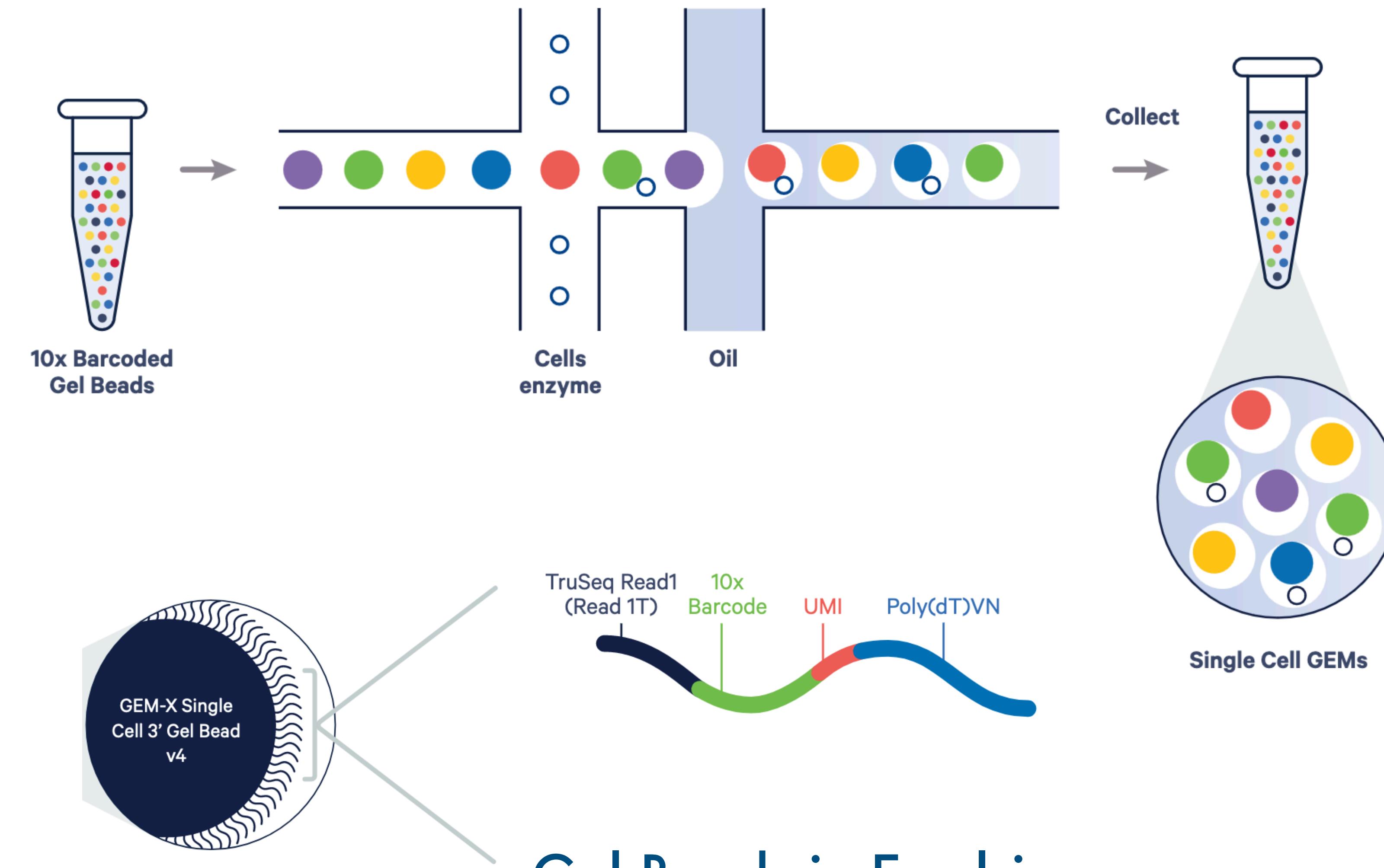
Mikrofluidik zur Erfassung einzelner Zellen

Bewegung winziger Flüssigkeitsmengen durch sehr enge Kanäle (mit einem Durchmesser von einigen zehn bis hundert Mikrometern)



Einzelzell-Transkriptomik, Springer 2024, Buch

10X Single Cell RNA Sequencing



10X Protokolle für die Einzelzell

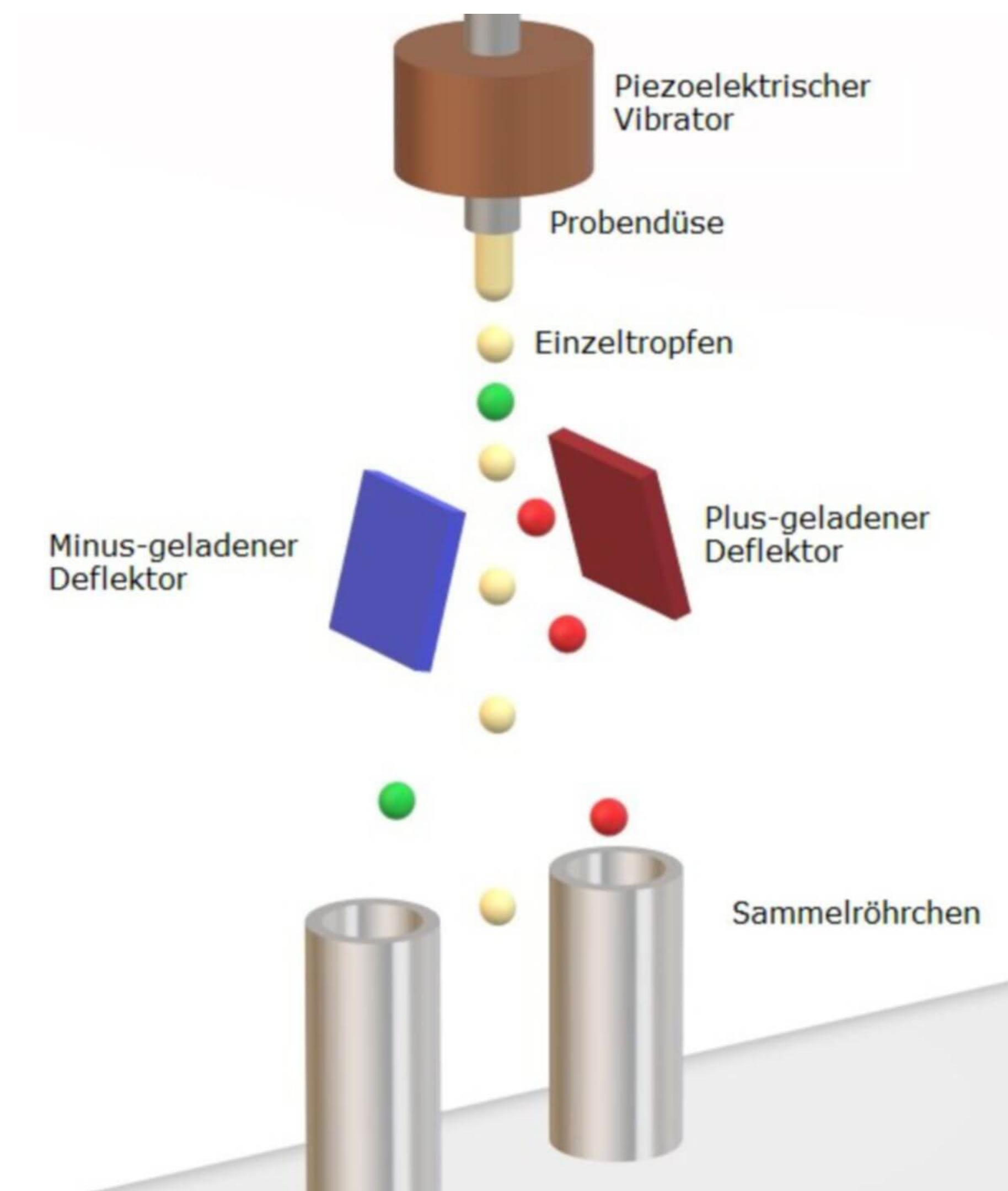
basieren auf Tröpfchen und verwenden eindeutige molekulare Identifikatoren (UMIs), um zu vermeiden, dass ein RNA-Fragment mehr als einmal gezählt wird.

10X hat den höchsten Durchsatz auf dem Markt und ist daher auch teuer.

Es liefert keine vollständigen Sequenzinformationen.

FACS zur Erfassung einzelner Zellen

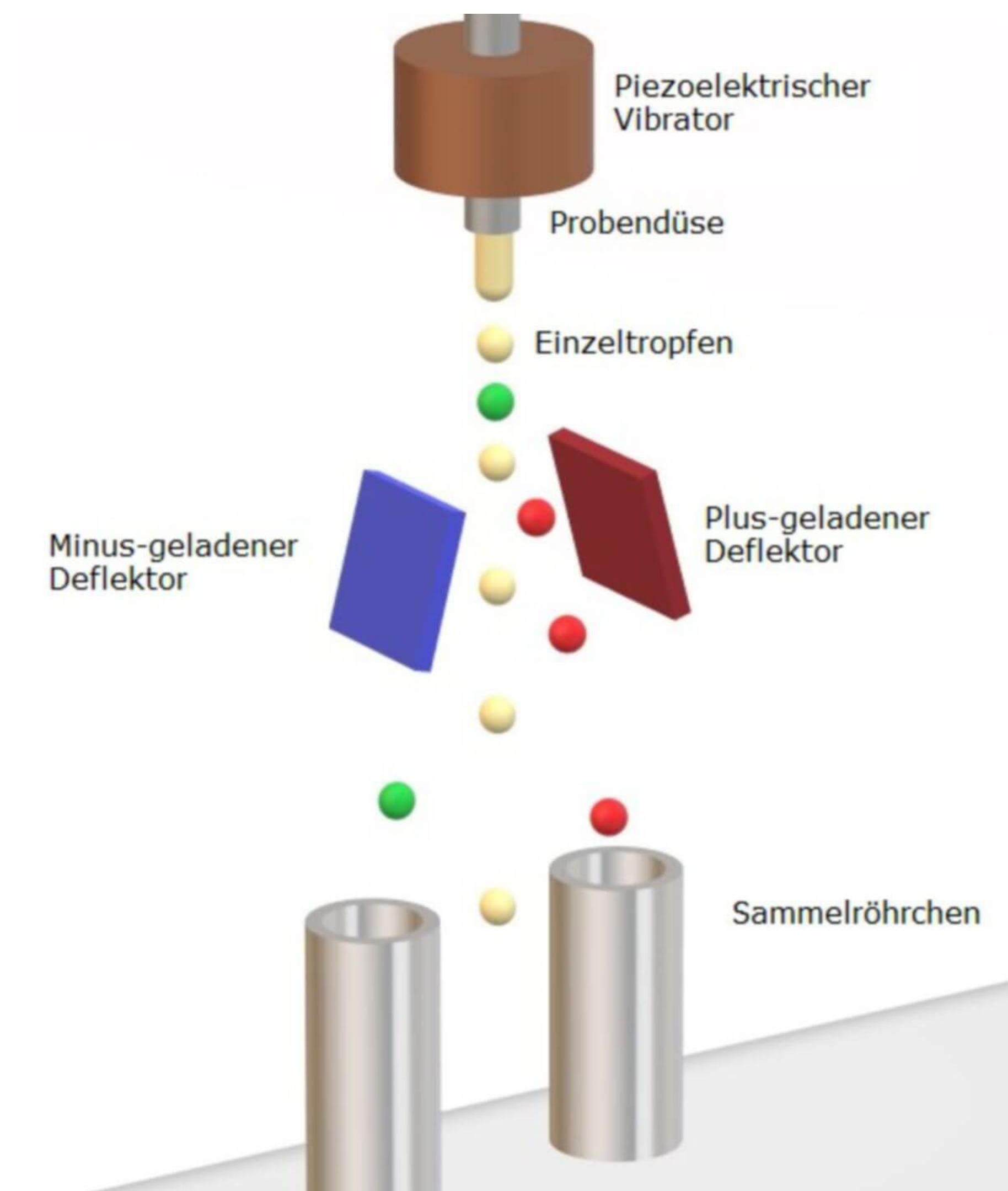
FACS, kurz für fluorescence-activated cell sorting, ist eine Analysemethode, bei der fluoreszenz-markierte Zellen abhängig von bestimmten Eigenschaften in unterschiedliche Probengefäße sortiert werden.



FACS zur Erfassung einzelner Zellen

Beispiel: Falls wir kennen, welche Krebszellen welche Oberflächenproteine sich besitzen, dann können wir diese Zellen als folgende sortieren:

1. ein stark grün fluoreszierender Farbstoff ist mit dem Antikörper von diesen Oberflächenproteinen gekoppelt.
2. Oberflächenproteinen verbinden zum Antikörper damit die Tumorzellen grün markiert.
3. Der Sorter lädt alle Flüssigkeitströpfchen, die eine grün leuchtende Zelle enthalten, positiv auf. Sie werden aus dem senkrechten Flüssigkeitsstrom abgelenkt und landen in einem Sammelröhrchen, während die nicht markierten Zellen verworfen werden.



Smart-Seq pipeline

Smart-seq2 ist ein beliebtes FACS-basiertes Protokoll für die EinzelzelleRNA-Sequenzierung.

Der Durchsatz nicht besonders hoch.

Sie können jedoch auswählen, welche Zellen Sie sequenzieren möchten.

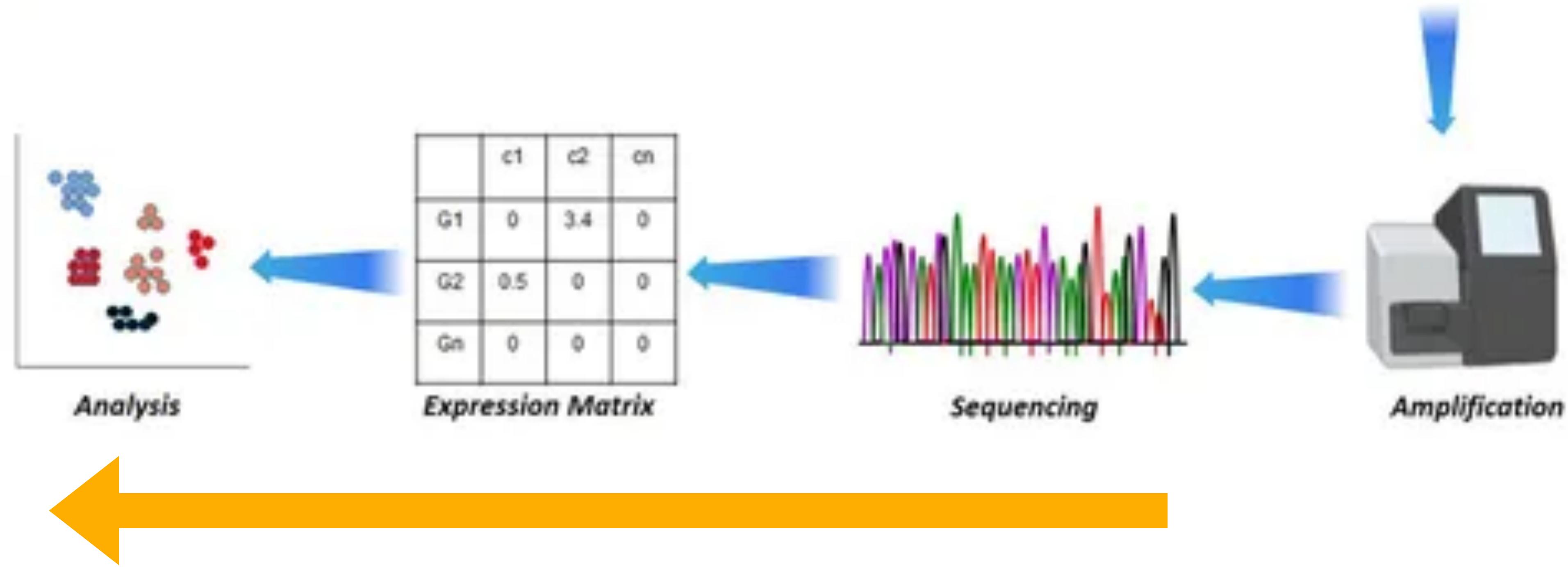
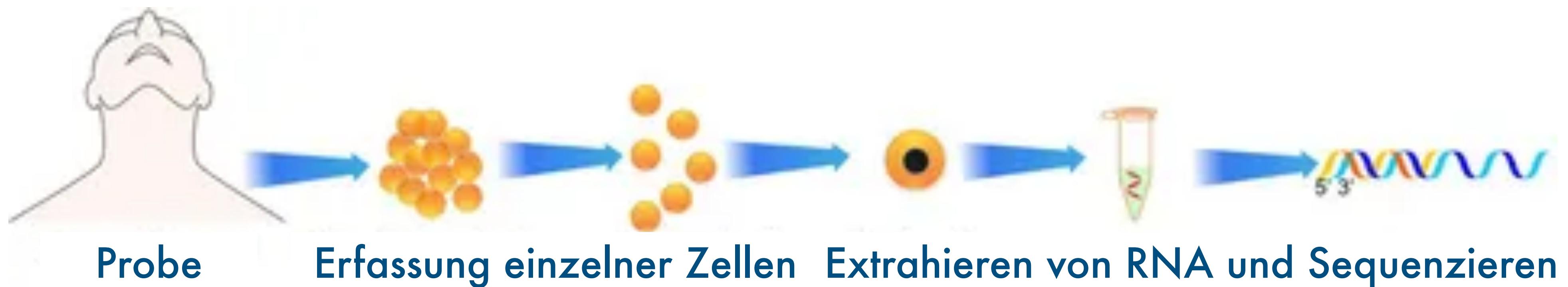
Mit diesen Protokollen erhalten Sie Informationen über das gesamte Transkriptom.

Smart-Seq vs 10X

In der Regel enthalten 10x-Daten im Vergleich zu Smart-Seq deutlich mehr Zellen, die mit geringerer Tiefe sequenziert wurden.

Smart-Seq2 bietet im Vergleich zum droplet-basierten 10X die höchste Empfindlichkeit, eine höhere Erfassungsfähigkeit und geringere Kosten.

Einzelzell-Transkriptomik



menschliches Genom

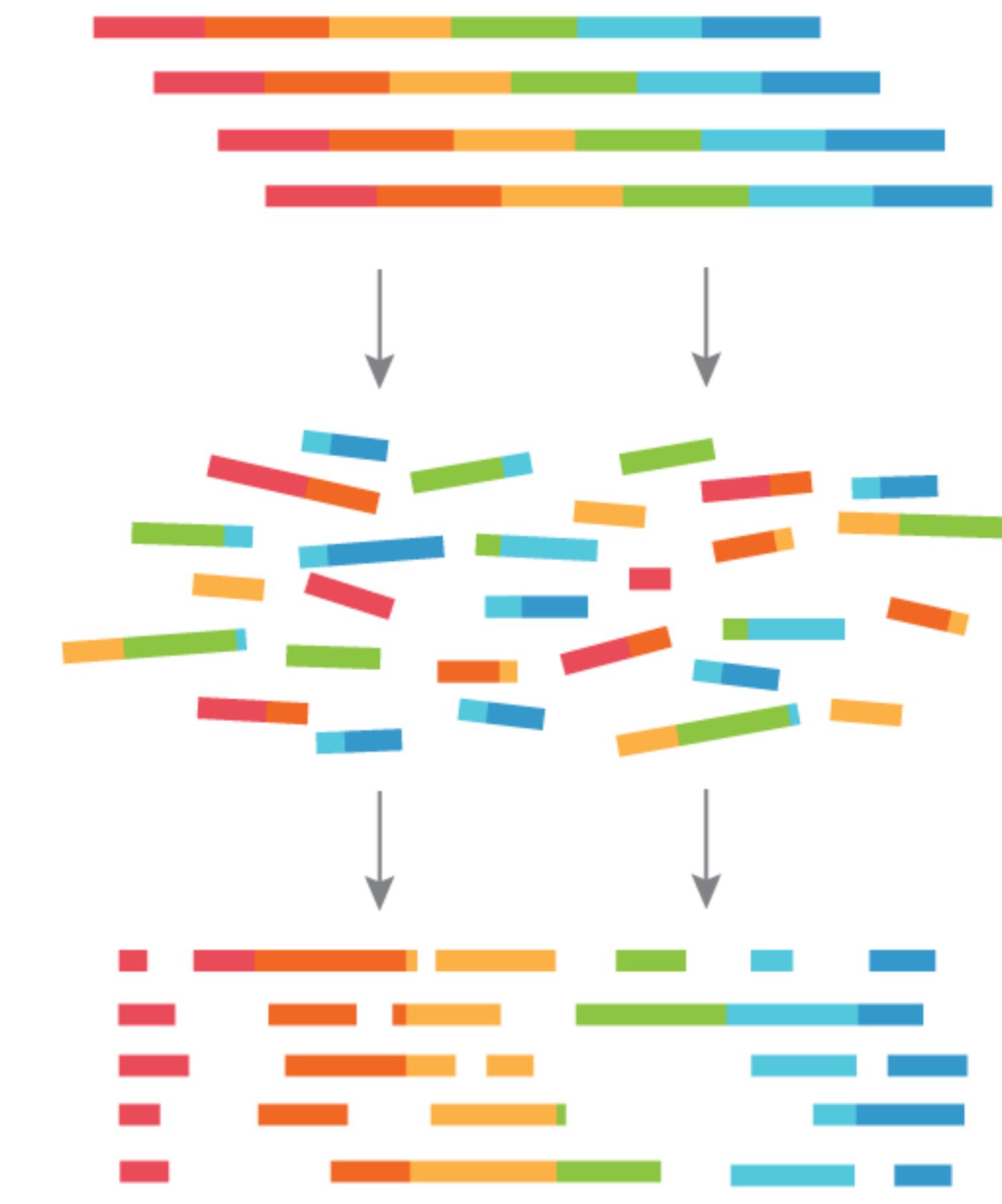
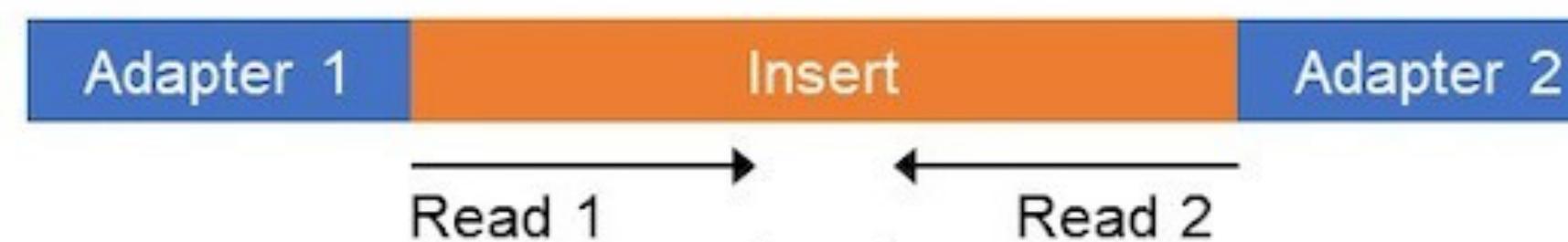
ist  Buchstaben lang,

davon kodieren  % für ein Protein

und es hat über  Gene.

Next Generation Sequencing

Wir müssen das Genom in kleine Abschnitte zerlegen und diese sequenzieren. Adapter an diesen Abschnitten helfen uns dabei, sie zu sequenzieren. Aber wir müssen diese Adapter entfernen. Dann gleichen wir die kurzen Abschnitte wieder mit einem Referenzgenom ab.



ATGTTCCGATTAGGAAACCTATCTGTAACGTGTTCAATTCAAGTAAAGGAGGAAA

frisch aus Sequenzer: Fastq-Dateien

Eine FASTQ-Datei hat vier Zeilen pro Sequenz:

Zeile 1 beginnt mit einem „@“-Zeichen, gefolgt von einem Titel oder ID.

Zeile 2 beinhaltet die Sequenzbuchstaben.

Zeile 3 beginnt mit einem „+“-Zeichen

Zeile 4 codiert die Qualitätskennzahlen für die Sequenz in Zeile 2 und muss die gleiche Anzahl an Symbolen enthalten wie Buchstaben in der Sequenz sind.

Eine FASTQ-Datei mit einer einzigen Sequenz kann folgendermassen aussehen:

```
@SEQ_ID
GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTT
+
! ''*( ( ( (***) ) %%%++ ) ( %%% ) . 1***-+*' ) )**55CCF>>>>>CCCCCCCC65
```

Qualitätskontrolle

Q in Fastq-Dateien steht für Qualitätswerte.

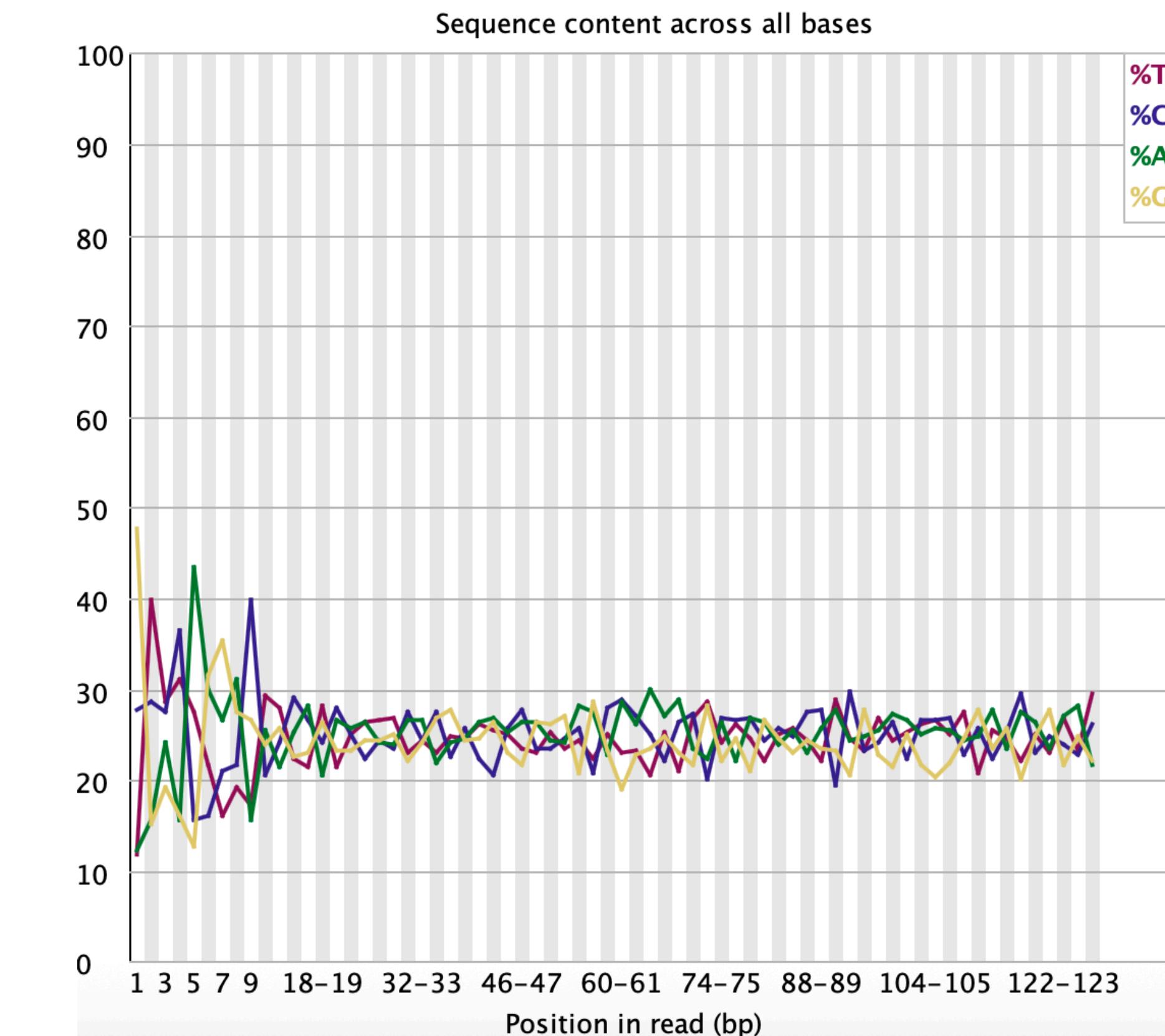
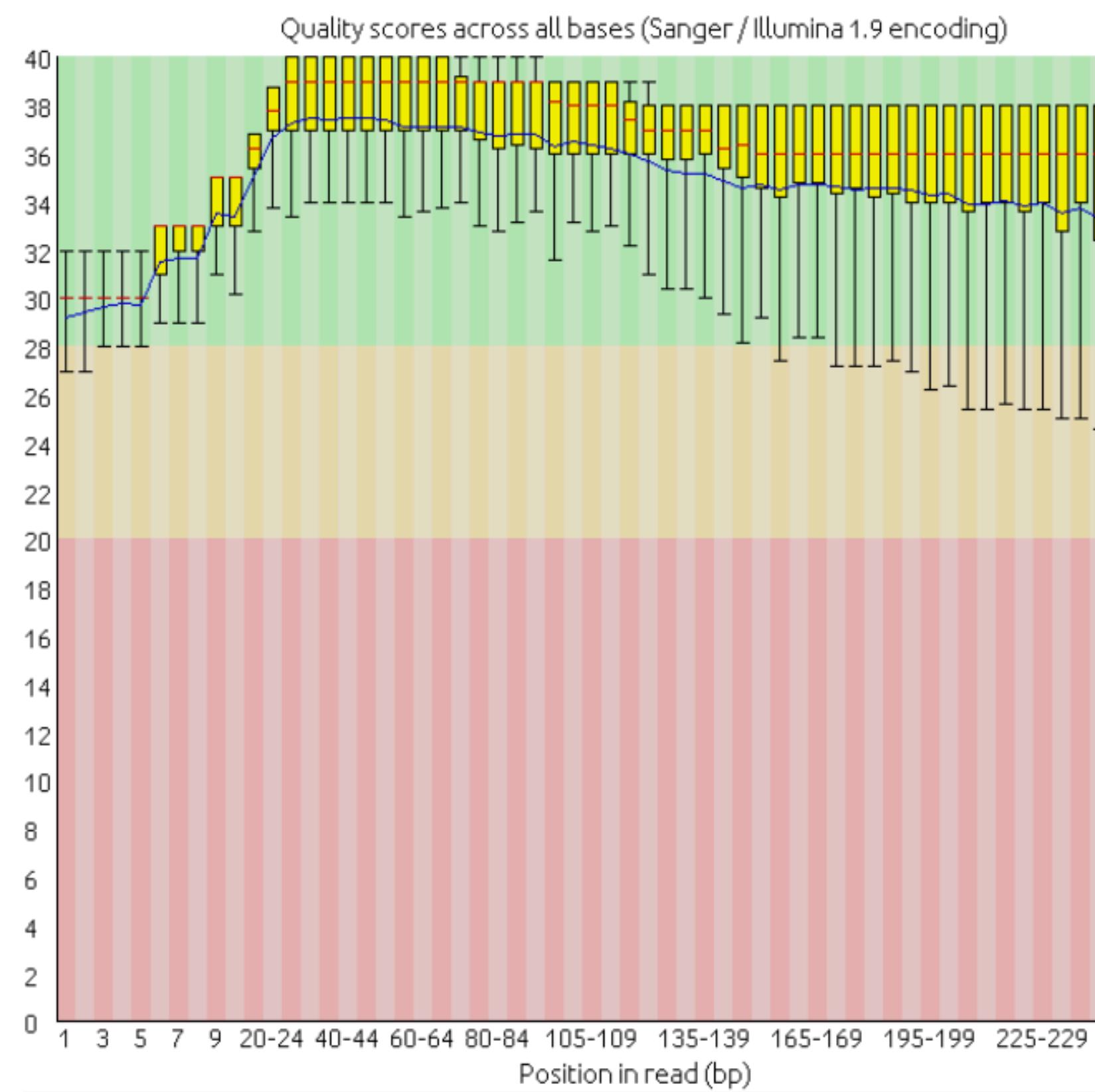
Jedes Zeichen steht für einen sogenannten Phred-Qualitätswert.

Phred Quality Score	Wahrscheinlichkeit einer falschen Basenbestimmung	Basenbestimmungsgenauigkeit
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

```
@SEQ_ID
GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTT
+
!'''*( ((***+) )%%%++ ) (%%% ) .1***-+*'') )**55CCF>>>>>CCCCCCCC65
```

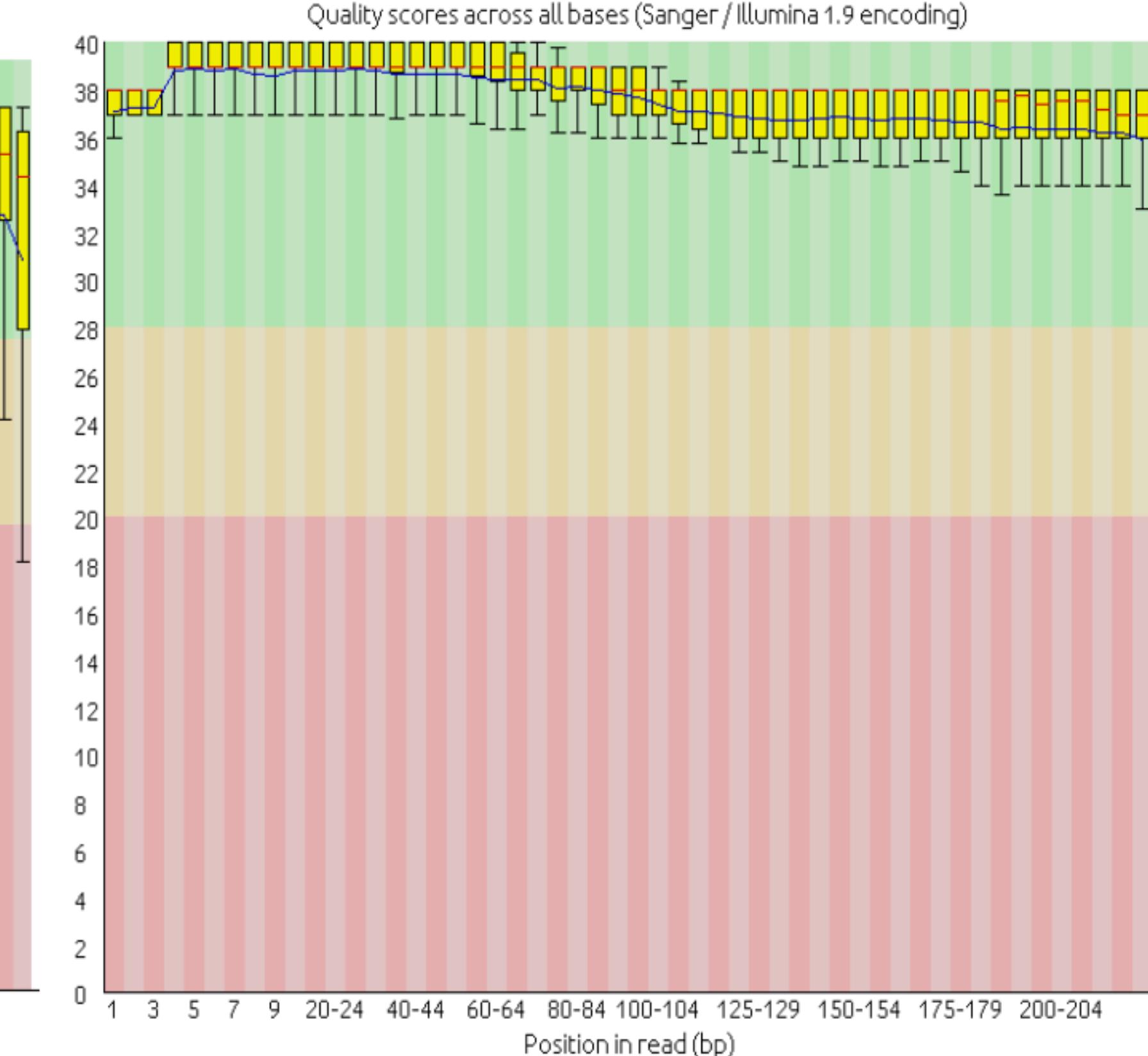
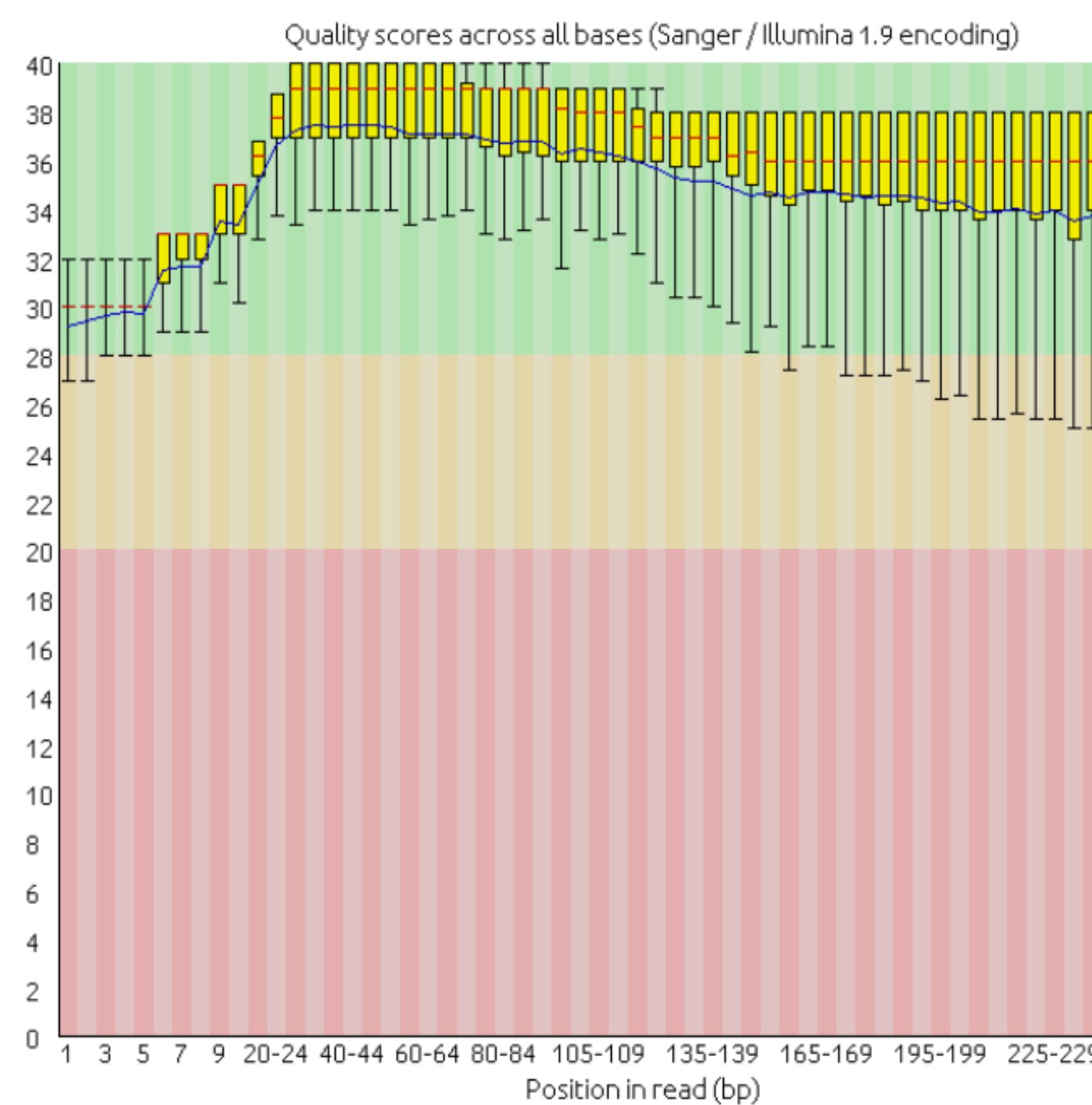
FastQC - App für Fastq-Dateien

Kurze Reads werden übereinander gestapelt.



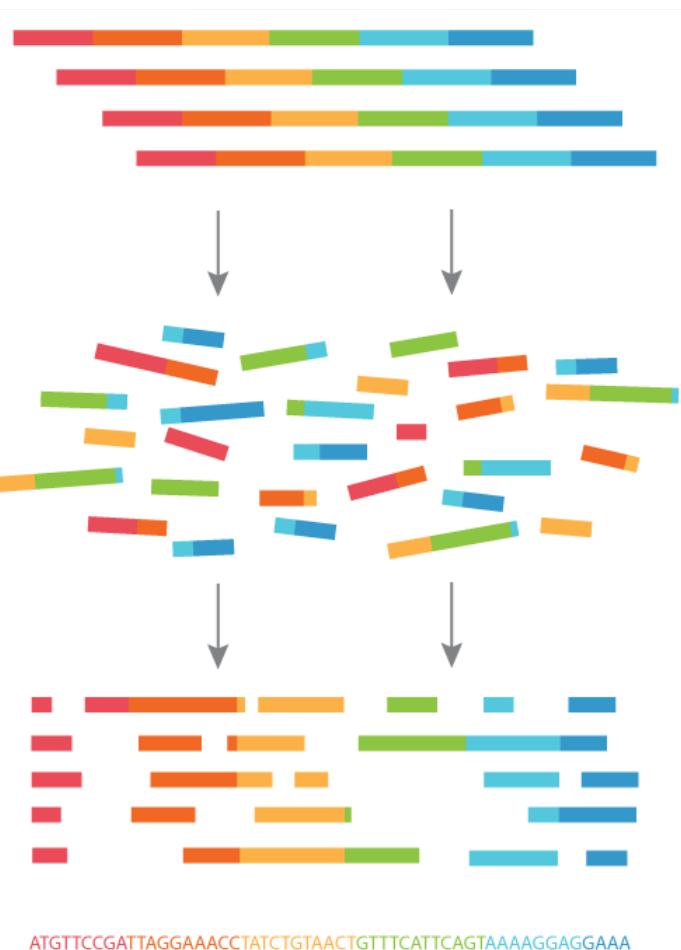
Datenbereinigung

FASTQ-Dateien werden getrimmt, um Sequenzdaten von geringer Qualität und Adaptersequenzen zu entfernen, was die Ergebnisse der nachgelagerten Analyse verbessert. Wir möchten Qualitätswerte > 30.

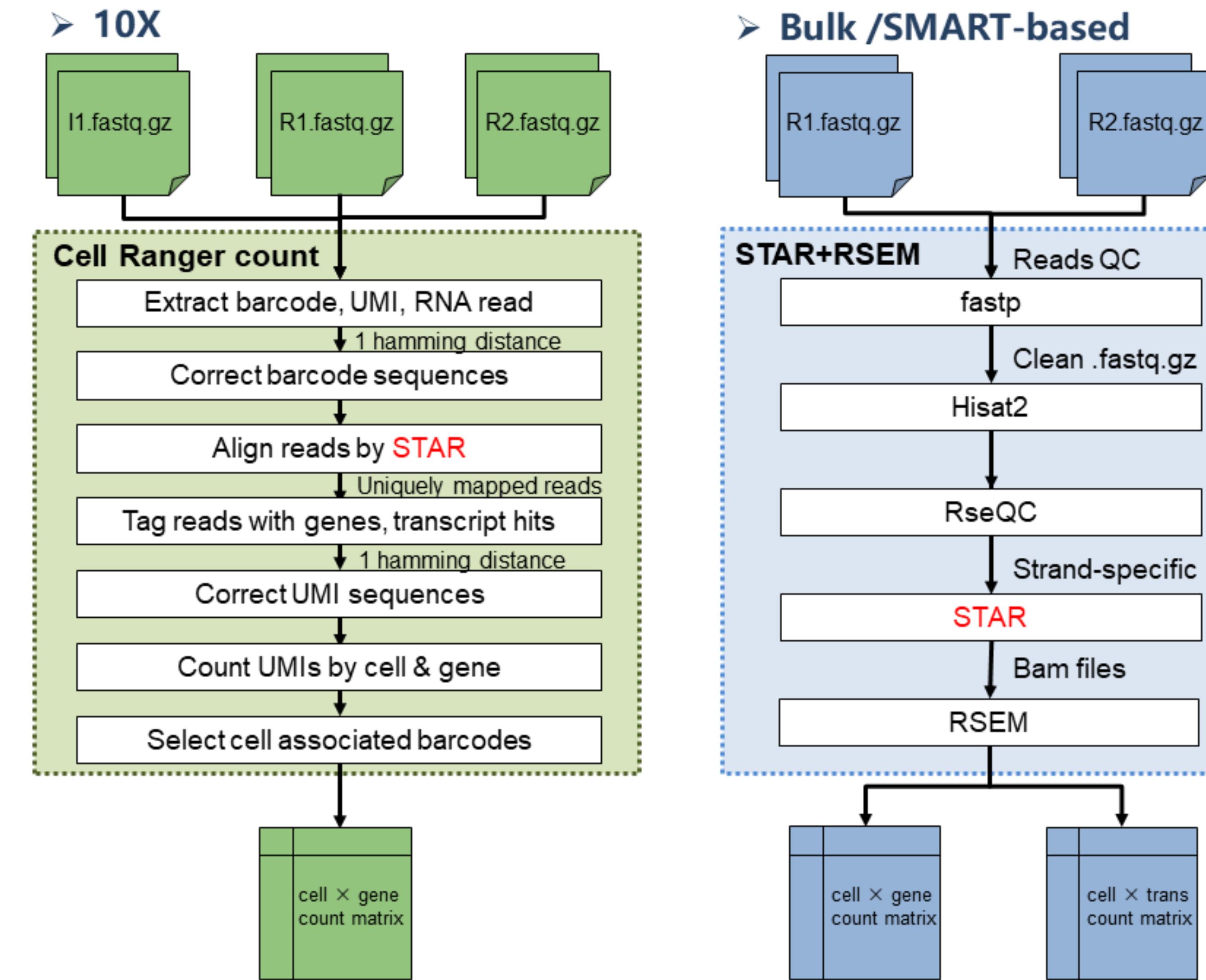


Alignment

8661 8671 8681 8691 8701 8711 8721 8731 8741 8751 8761 8771
901TCCCACTCTCAGAACACTGAGAAAAGTGAGGCATGGGTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTACAAGACTGGTGAGGGAAAGGTGTAACCTGTTGTCA
M.....
AGCTCCCAC TCTCAGAACACTG tgggtttctgggctggtacaggagctcgatgtgcttctctacaagactggtgaggaaagggtgtaacctgtttg
AGCTCCCAC TCTCAGCACACTG GTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTACAAGACTGGT GAGGGAAAGGTGTAACCTGTTGTCA
AGCTCCCAC TCTCAGAACACTG GTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTACAAGACTGGTAAGGGAAAGGTGTAACCTGTTGTCA
AGCTCCCAC TCTCAGAACACTG GTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTACAAGACTGGAGAGGGAAAGGTGTAACCTGTTGTCA
AGCTCCCAC TCTCAGCACACTG GTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTACAAGACTGGT GAGGGAAAGGTGTAACCTGTTGTCA
AGCTCCCAC TCTCAGAACACTG GTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTACAAGACTGGTAAGGGAAAGGTGTAACCTGTTGTCA
AGCTCCCAC TCTCAGAACACTGAGAAAAGTGAGGC A GTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTACAAGACTGGAGAGGGAAAGGTGTAACCTGTTGTCA
agctcccactctcagaacactgagaaaagtgaggcatggg tttctgg CGATGTGCTTCTCTACAAGACTGGT GAGGGAAAGGTGTAACCTGTTGTCA
agctcccactctcagaacactgagaaaagtgaggcatggg tttctgg tataacctattgtcaga
agctcccactctcagaacactgagaaaagtgaggcatggg tttctgg TAACCTGTTGTCA
agctcccactctcagaacactgagaaaagtgaggcatggg tttctgg GTTGTCA
agctcccactctcagaacactgagaaaagtgaggcatggg tttctgg GTTGTCA
agctcccactctcagaacactgagaaaagtgaggcatggg tttctgg GTTGTCA
AACTGAGAAAAGTGAGGCATGGGTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTACAAGACTGGT GAGG GTTGTCA
AACTGAGAAAAGTGAGGCATGGGTTATGGGATGGTACAGGAGCTCGATGTGCTTCTCTACAAGACTGGT GAGG GTTGTCA
AACTGAGAAAAGTGAGGCATGGGTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTACAAGACTGGT GAGG GTTGTCA
AACTGAGAAAAGTGAGGCATGGGTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTACAAGACTGGT GAGG GTTGTCA
AACTGAGAAAAGTGAGGCATGGGTTATGGGATGGTACAGGAGCTCGATGTGCTTCTCTACAAGACTGGT GAGG GTTGTCA
AACTGAGAAAAGTGAGGCATGGGTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTACAAGACTGGT GAGG GTTGTCA
AACTGAGAAAAGTGAGGCATGGGTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTACAAGACTGGT GAGG GTTGTCA
AACTGAGAAAAGTGAGGCATGGGTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTACAAGACTGGT GAGG GTTGTCA
GT TGTCA



Einzelzell-Transkriptomik Analyse: Erstellen der Matrix

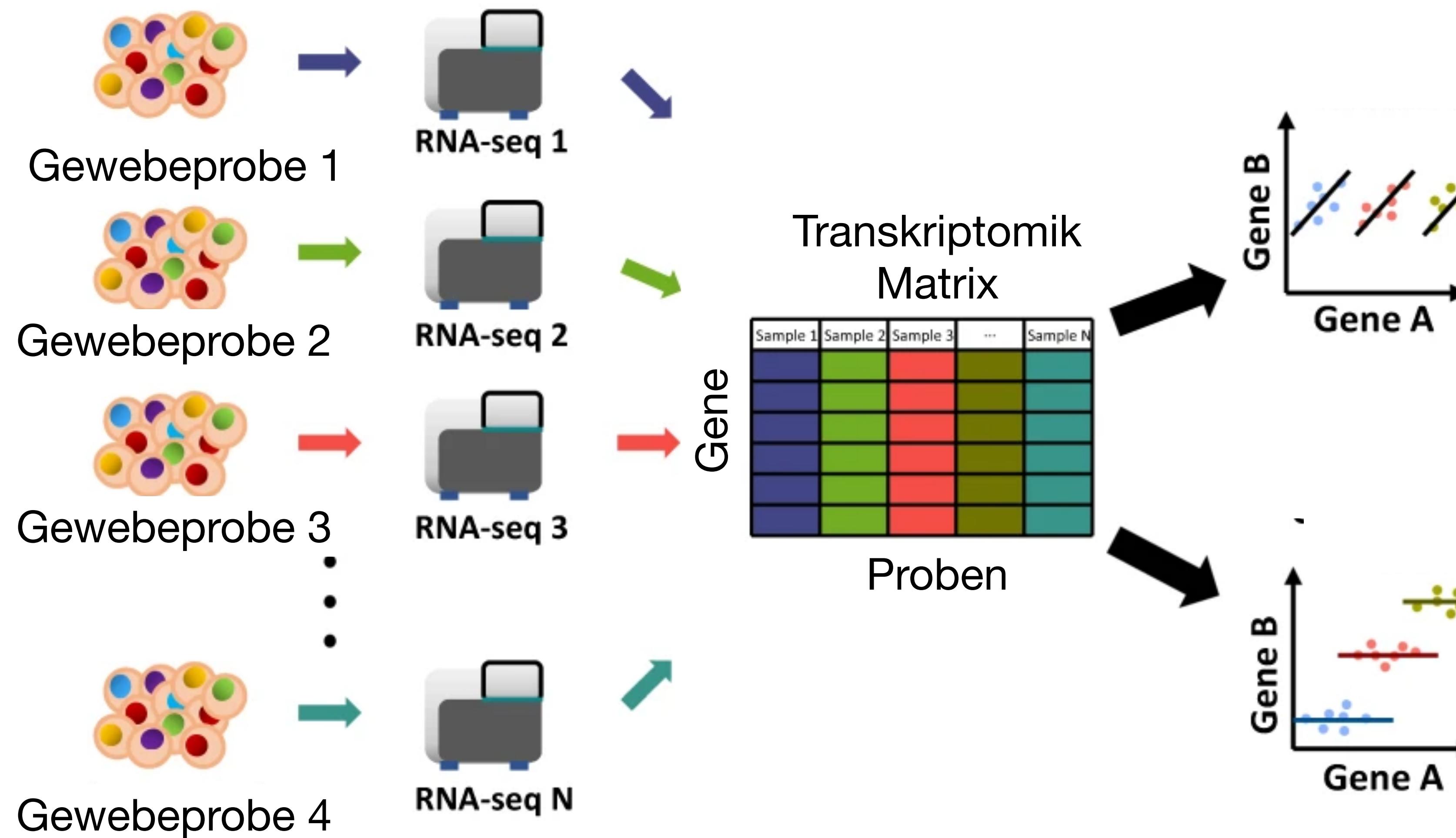


Einzelzell-Transkriptomik-Matrix

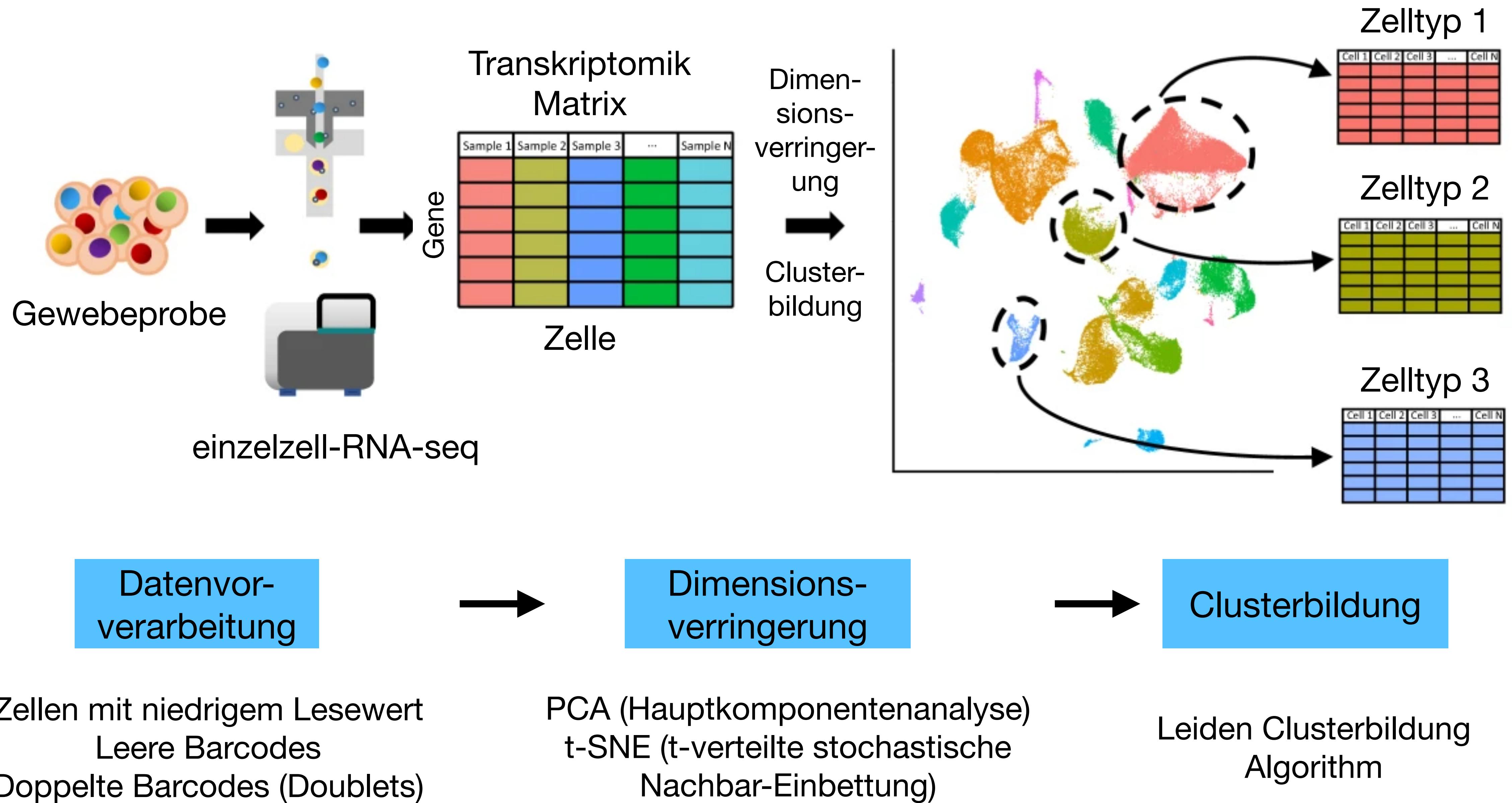
Tabelle, die die Genexpressionsdaten einzelner Zellen darstellt. Jede Zeile der Matrix entspricht einer bestimmten Zelle, und jede Spalte steht für ein Gen. Der Wert an jeder Schnittstelle oder Zelle gibt die Anzahl der für dieses Gen in dieser Zelle nachgewiesenen Transkripte an und liefert so einen Überblick über das molekulare Profil jeder Zelle.

	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...
...
...
GeneM	25	0	.	0

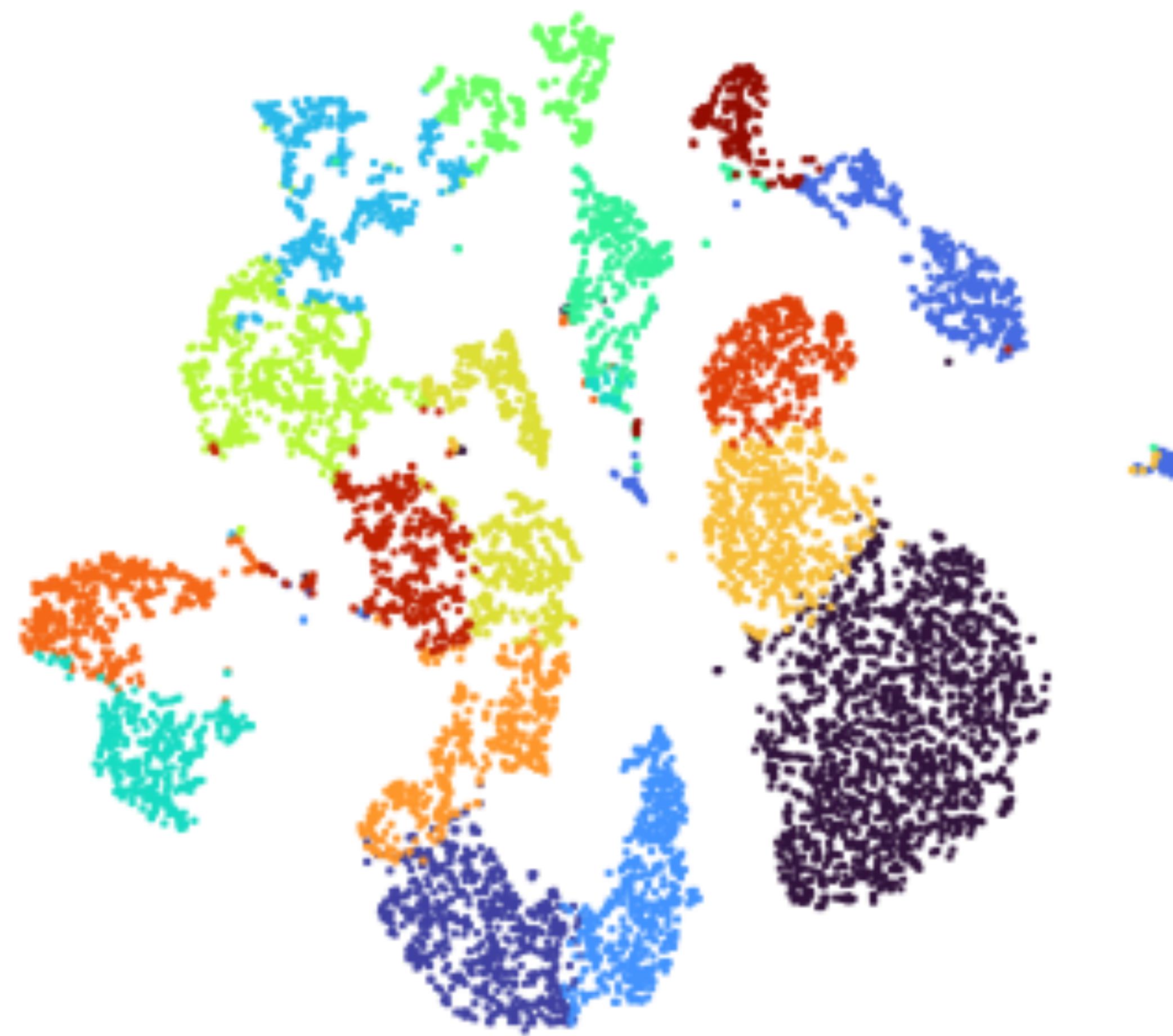
RNA-Sequenzierungsanalyse



Einzelzell-RNA-Sequenzierungsanalyse



Einzelzell-RNA-Sequenzierungsanalyse Resultate



Zelltype



Leukemia Patient

Gesund

Einzelzell-RNA-Sequenzierungsanalyse

R Analyse

- Seurat Paket
- Kann nur verarbeitete Daten lesen
(zB 10X)

Python Analyse

- Scvi und Scanpy
- Kombiniert mit maschinellem Lernen (scikit)

Cell-Ranger/Loupe Browser

- 10X Omik Datensätze
- User Friendly
- Standard-Datenanalyse und -Visualisierung

heutiges Training in R

Für dieses Tutorial analysieren wir einen Datensatz von peripheren mononukleären Blutzellen (PBMC), der von 10X Genomics frei verfügbar ist. Es handelt sich um 2.700 Einzelzellen, die auf dem Illumina NextSeq 500 sequenziert wurden. Die Rohdaten und R-Skript sind online.

```
## 3 x 30 sparse Matrix of class "dgCMatrix"
##
## CD3D  4 . 10 . . 1 2 3 1 . . 2 7 1 . . 1 3 . 2  3 . . . . 3 4 1 5
## TCL1A . . . . . . . 1 . . . . . . . . . . 1 . . . . . . .
## MS4A1 . 6 . . . . . 1 1 1 . . . . . . . 36 1 2 . . 2 . . .
```

1. Datenvorverarbeitung

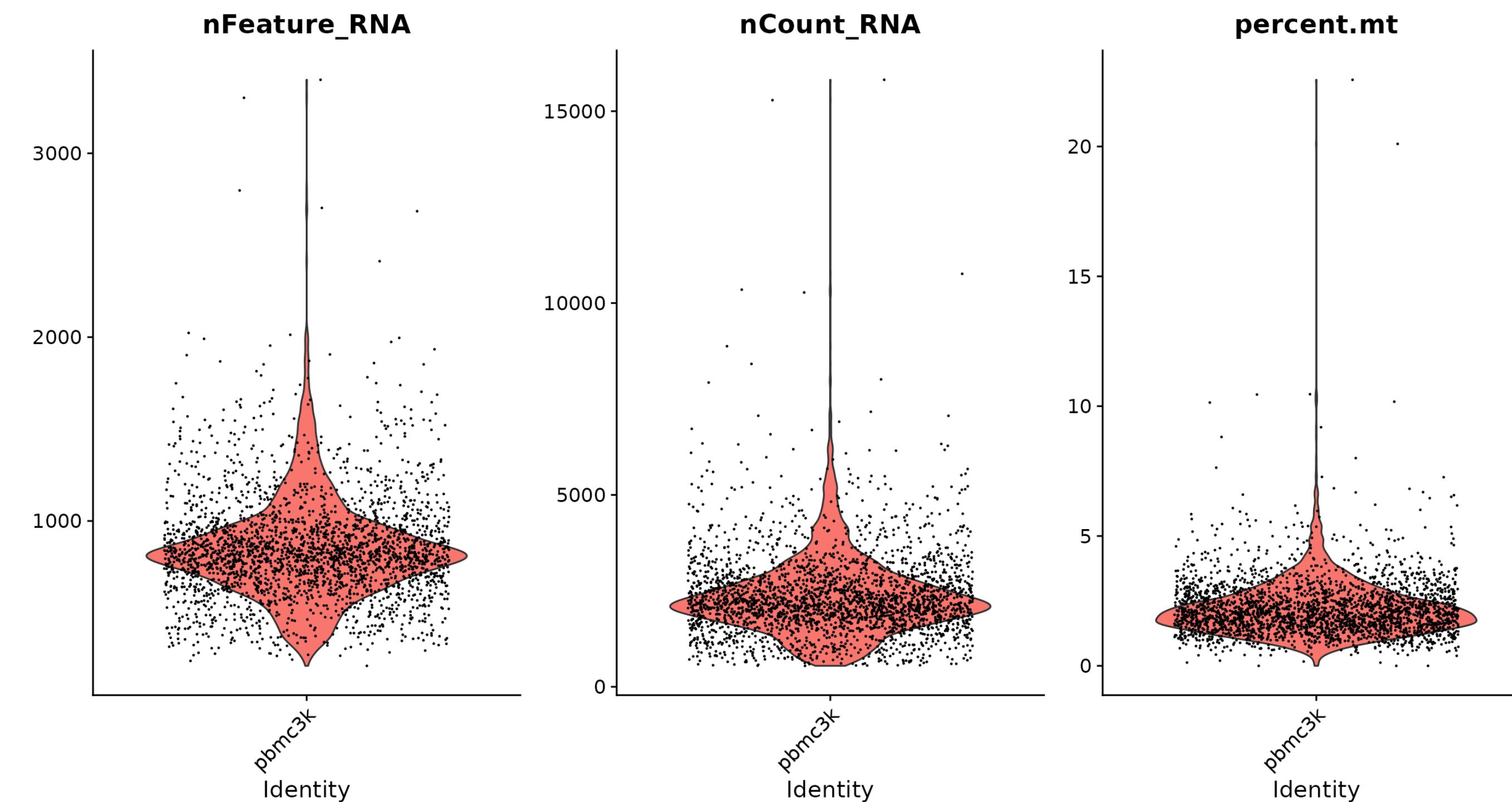
Zellen von geringer Qualität oder leere Tröpfchen weisen oft nur sehr wenige Gene auf.

Zell-Doubletten oder -Multipletts können eine ungewöhnlich hohe Genanzahl aufweisen.

Zellen von geringer Qualität/sterbende Zellen weisen oft eine starke mitochondriale Kontamination auf.

1. Datenvorverarbeitung

Wir filtern Zellen, die mehr als 2.500 oder weniger als 200 einzigartige Merkmale aufweisen, sowie Zellen, die mehr als 5 % Mitochondrien enthalten.



Das Streben nach reduzierten Dimensionen: PCA

Das „Schülerleistungs“-Beispiel:

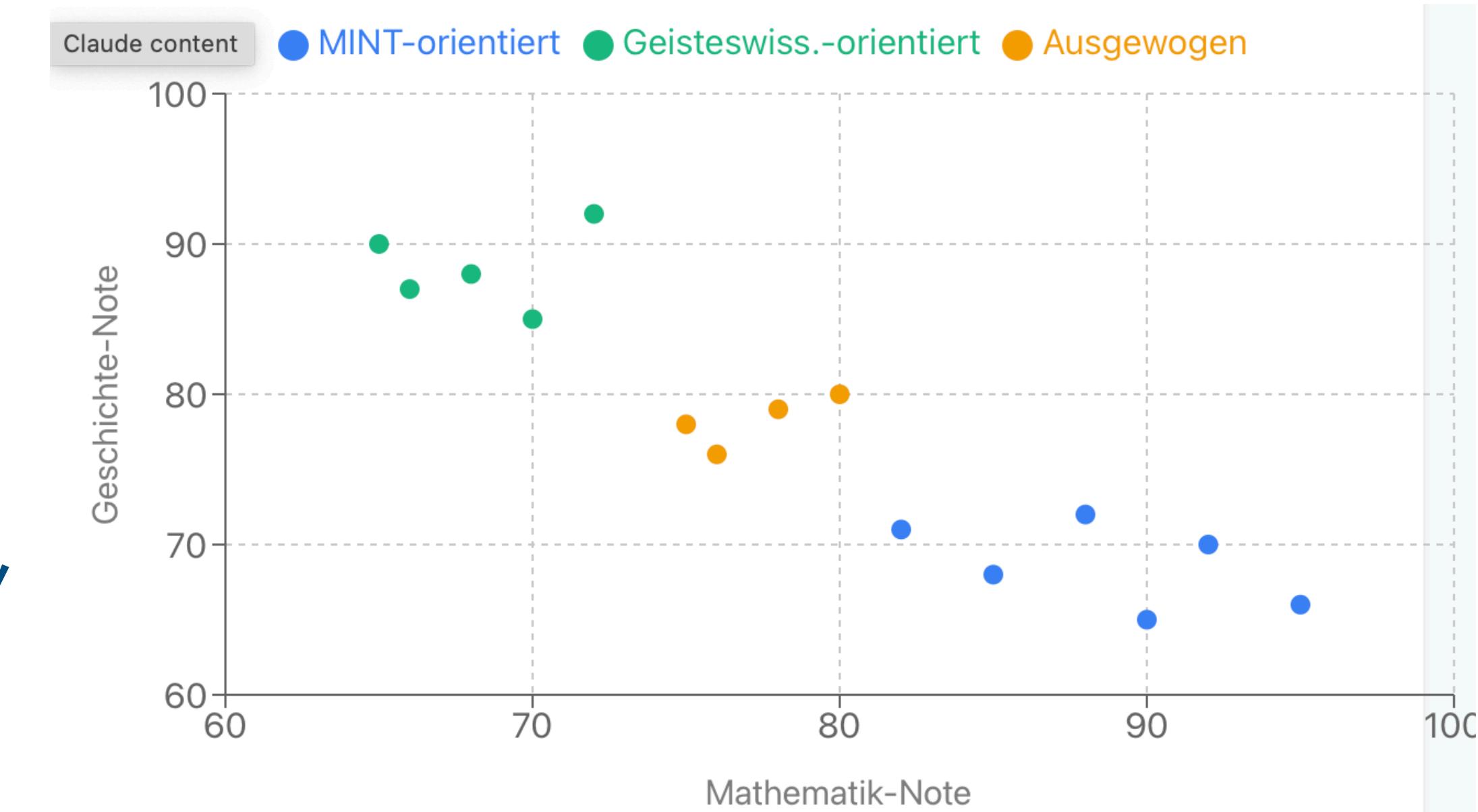
Stellt euch vor, ihr befragt 100 Schüler und erfasst ihre Noten in 6 Fächern: Mathematik, Physik, Chemie, Biologie, Geschichte und Englisch.

Wenn ihr die Daten betrachtet, bemerkt ihr etwas Interessantes:

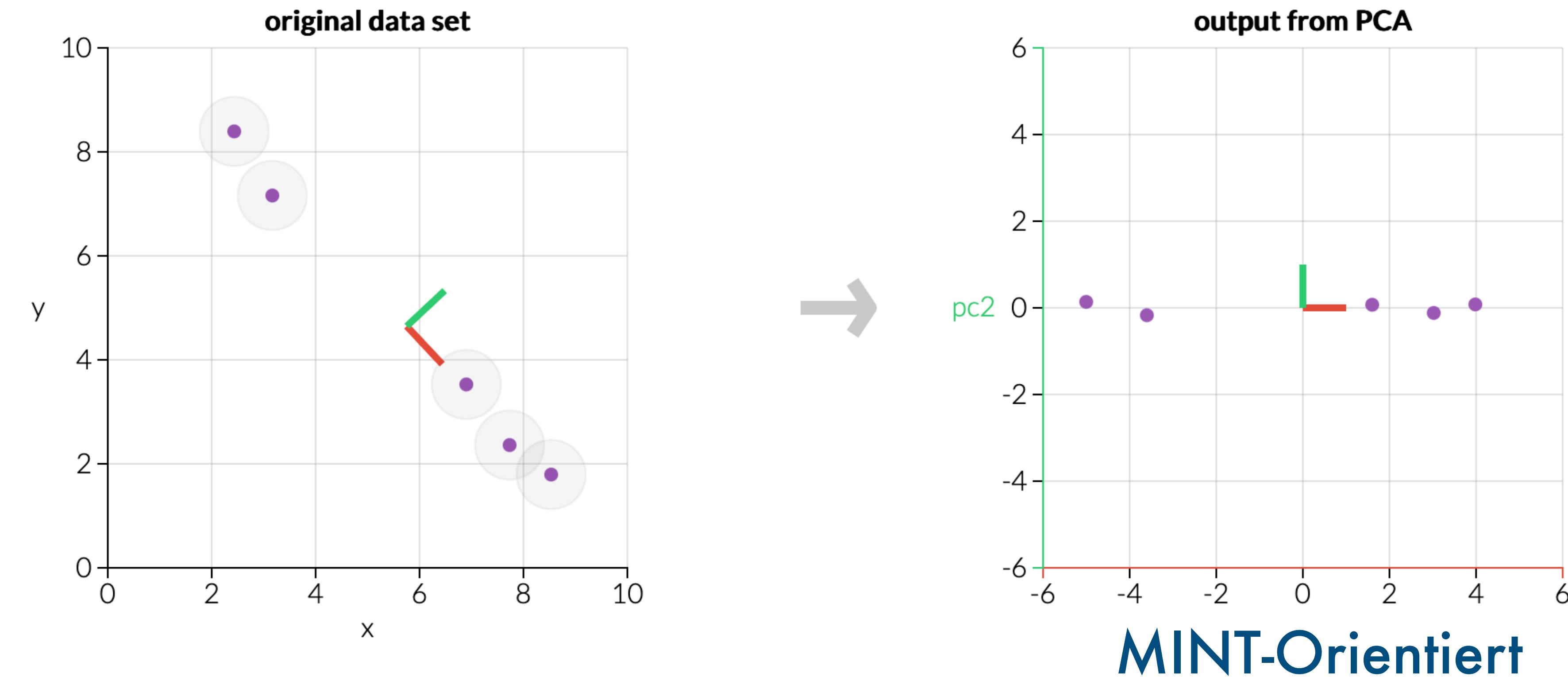
Schüler, die gut in Mathe sind, tendieren auch dazu, gut in Physik, Bio und Chemie zu sein

Schüler, die gut in Geschichte sind, tendieren dazu, gut in Englisch zu sein.

Hier kommt Dimensionsverringerung ins Spiel:
Anstatt 5 separate Noten zu verfolgen, können wir eine „MINT-Orientiert“-Achse oder eine „Geisteswissenschaften“-Achse definieren.

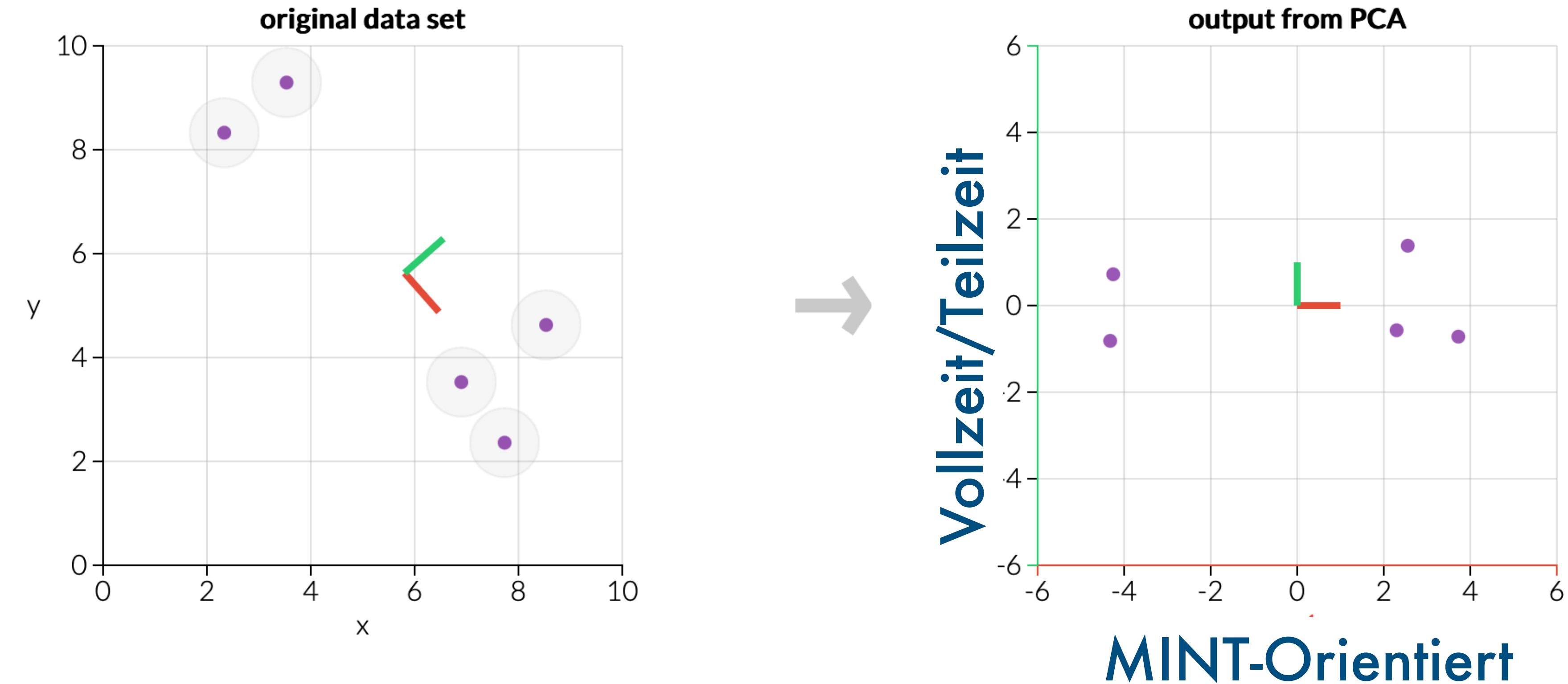


Hauptkomponentenanalyse (PCA)



zur Simulation: <https://setosa.io/ev/principal-component-analysis/>

Dimensionsverringerung durch PCA



zwei Dimensionen erklären die Daten: 2 Dimensionen statt 6 Fächer

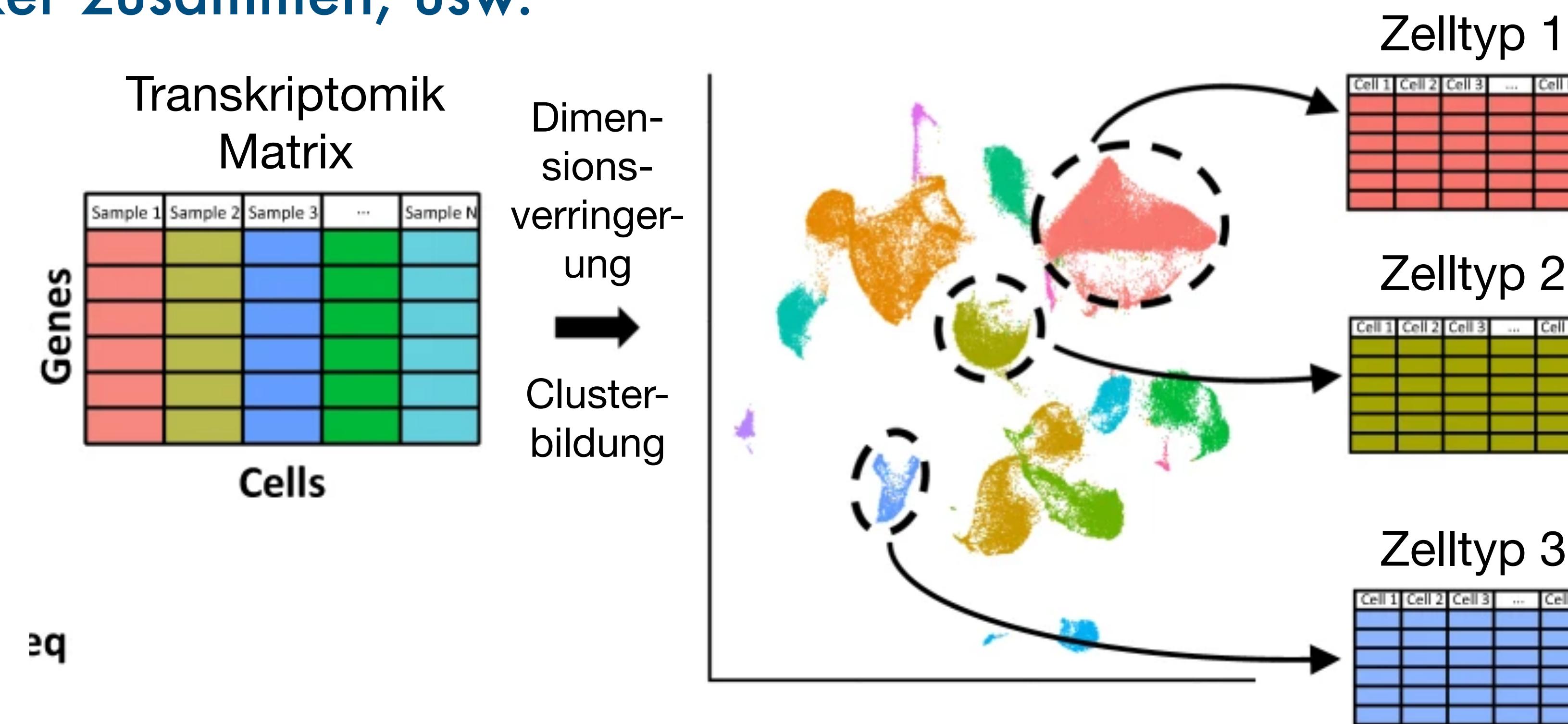
Zusammenfassung: PCA

Das Verfahren ist der Idee nach:

- Finde Achse mit grösster Datenvarianz (grösste Varianz der Daten, wenn man auf Achse projiziert) => 1. Hauptkomponenten (PC1)
- Projizierte Daten auf Datenraum der orthogonal = senkrecht zu PC1 steht und wiederhole bis du im 2D Fall angelangt bist
- 2-dimensionale PCA-Plots erlauben einem eine erste explorative Datenanalyse von hochdimensionalen Daten.
- Man sieht (potentiell), welche Datenelemente ähnlich sind und ob es Gruppenstrukturen (Cluster) gibt.
- Man erhält einen Hinweis, wie schwierig es für Machine Learning Methoden ist, Klassen zu bilden

Warum PCA in Transkriptomik?

Bei der Genexpression haben wir statt 5 Fächern 20.000 Gene. Statt 100 Schülern haben wir Tausende von Zellen. Gene, die zusammen variieren (wie Zelltyp-Marker), werden zur gleichen Hauptkomponente beitragen. - z.B. alle Immunzell-Marker zusammen, alle Neuronen-Marker zusammen, usw.

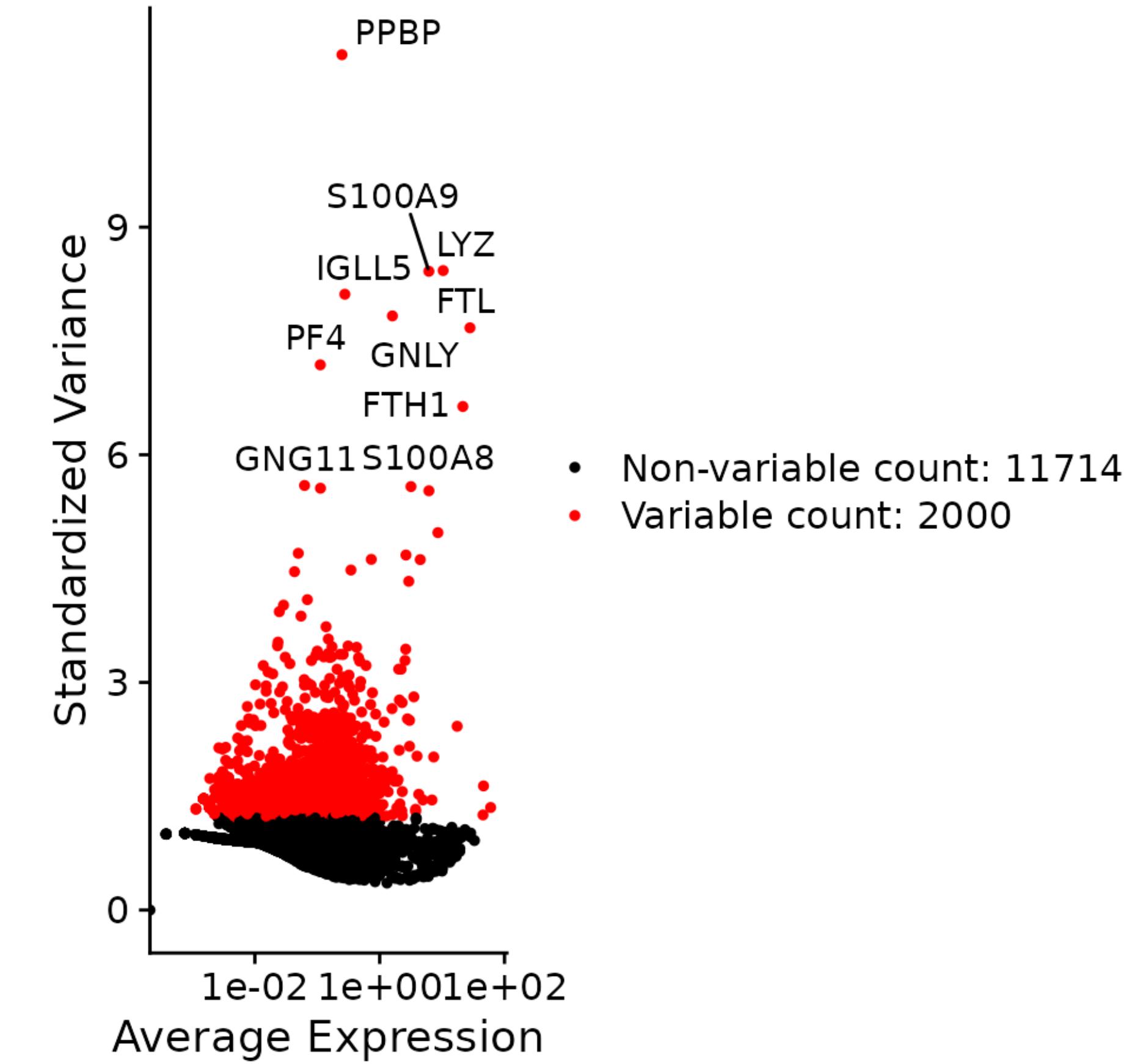


Warum PCA in Transkriptomik?

- Wenn Sie scRNA-seq-Daten messen, versuchen Sie, jede Zelle anhand der Expression ihrer Gene zu charakterisieren. Wenn ein Datensatz die Expression von 10.000 Genen misst, können Sie ihn als 10.000-dimensional beschreiben, was sehr viel ist.
- Obwohl diese Daten hochdimensional sind, kann ihre effektive Dimensionalität (d. h. die Anzahl der Dimensionen, die tatsächlich die Variation im Datensatz erfassen) viel geringer sein. Selbst wenn Sie beispielsweise über die Expressionsdaten für 10.000 Gene verfügen, können Sie die Unterschiede zwischen einer begrenzten Anzahl von Zelltypen möglicherweise mit nur wenigen Dimensionen quantifizieren.

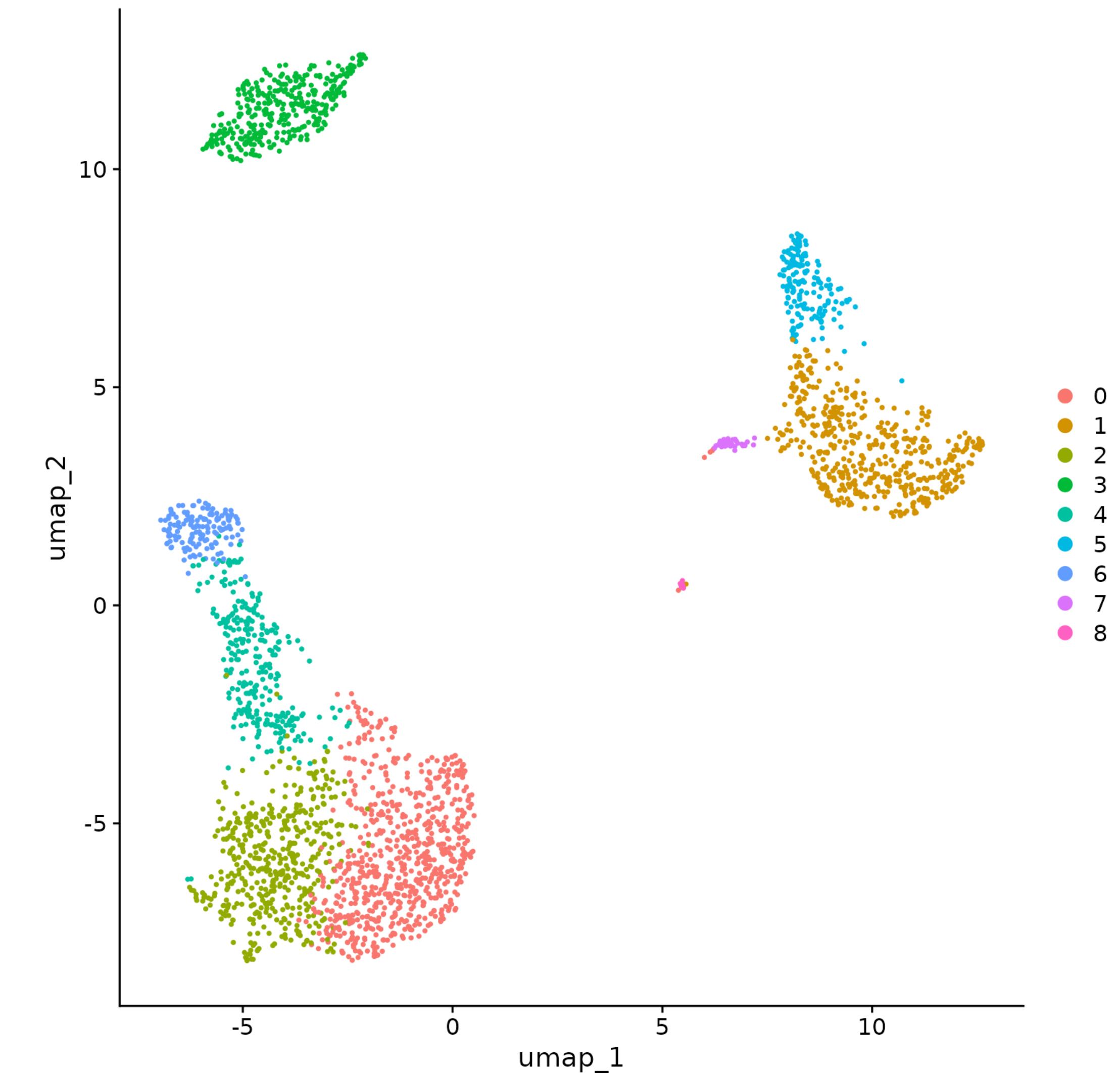
2. Dimensionsverringerung

Wir führen deswegen eine PCA für die skalierten Daten durch. Diese Methode reduziert die Dimensionalität der Daten, indem sie sie in einen Raum mit geringerer Dimension transformiert und dabei so viel Variation wie möglich beibehält. Sie sucht nach einer Kombination von Genexpressionen, den sogenannten Hauptkomponenten, die die grösste Variation im Datensatz erfassen.



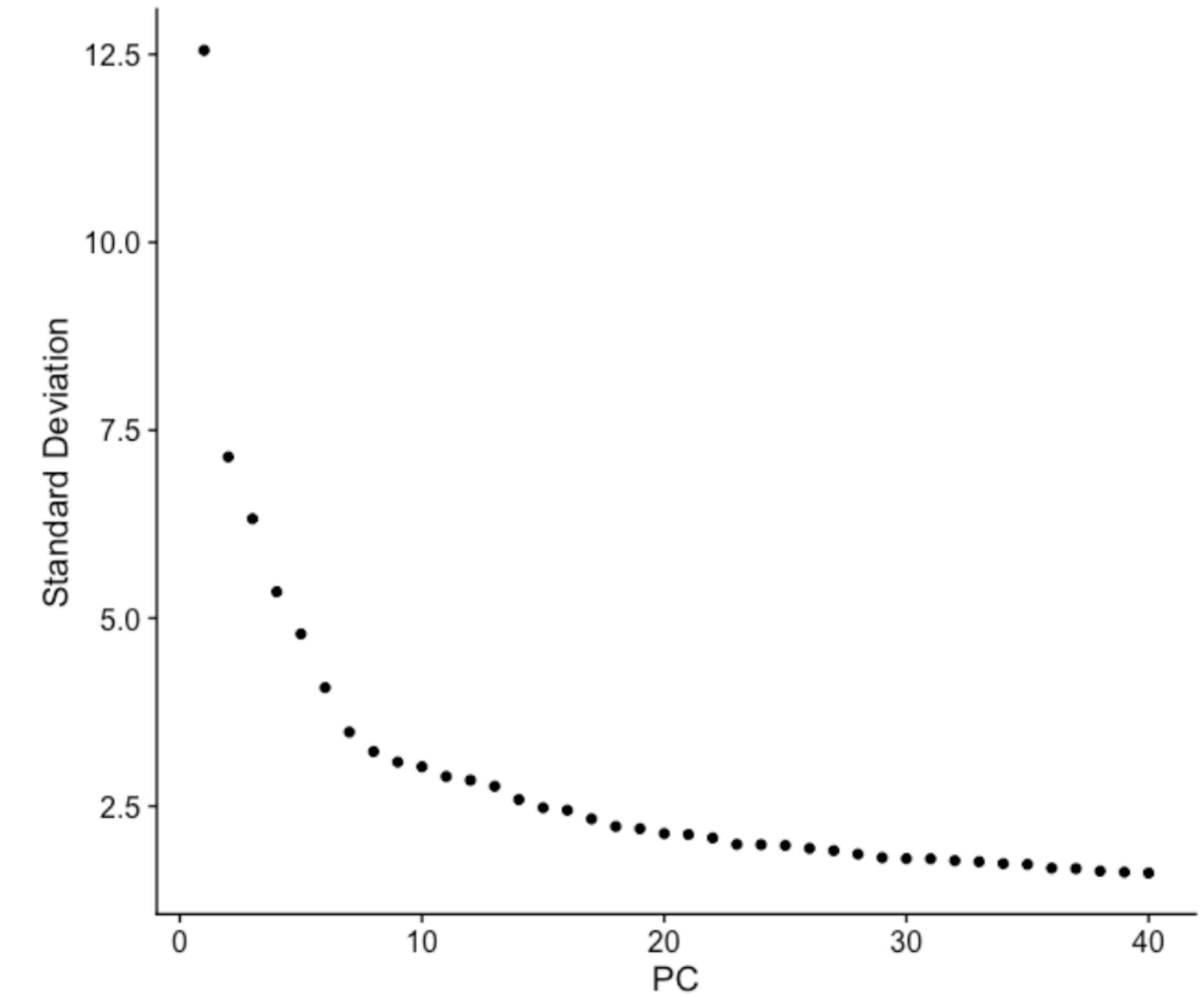
3. Clusterbildung von Zellen

Seurat führt zunächst eine graphbasierte Clusterbildung durch und führt anschliessend eine nichtlineare Dimensionsreduktion (UMAP/tSNE) durch. Dadurch werden ähnliche Zellen anhand ihrer Genexpression gruppiert, um Gruppen von Zellen zu finden, die häufiger miteinander in Verbindung stehen als andere.



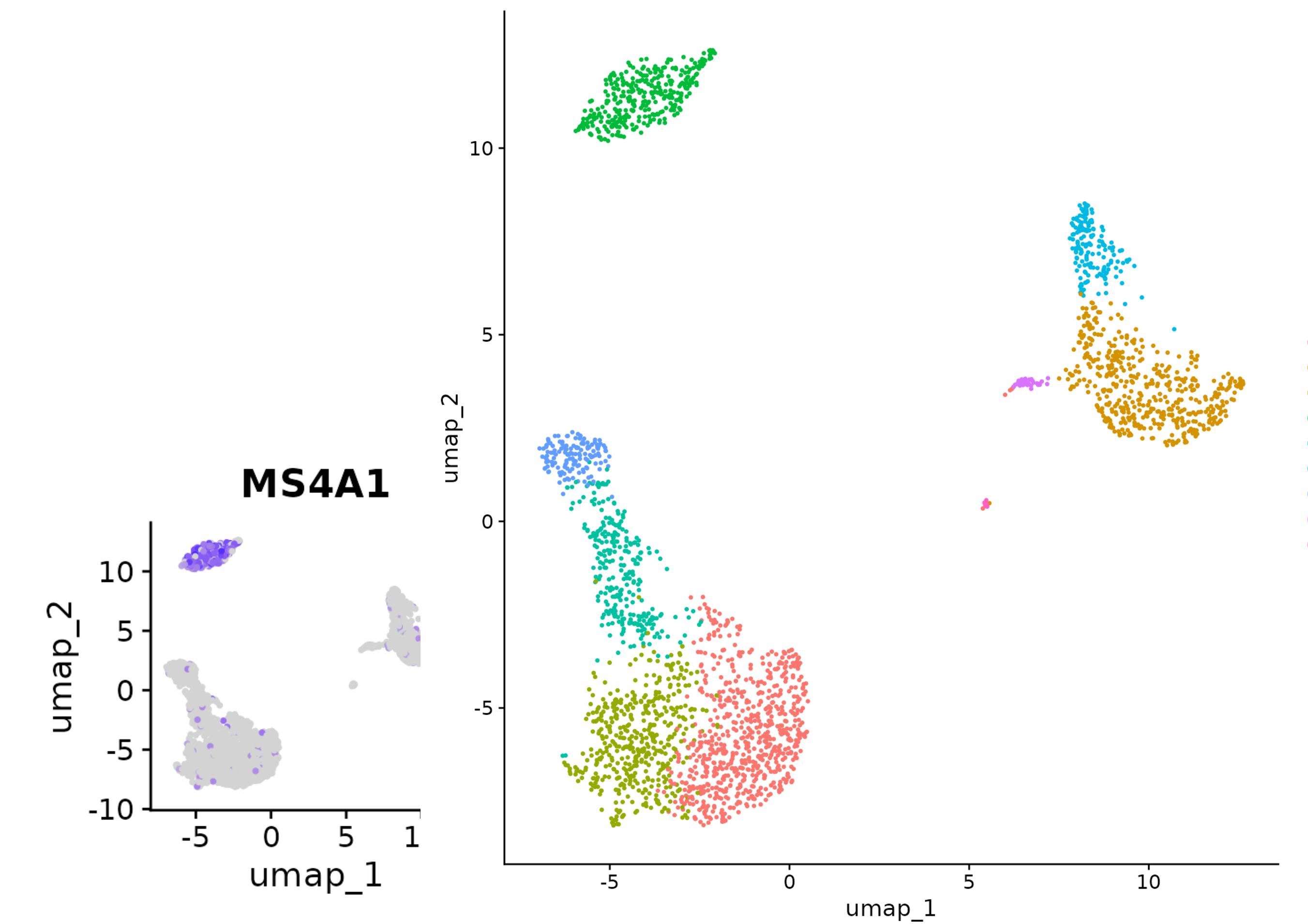
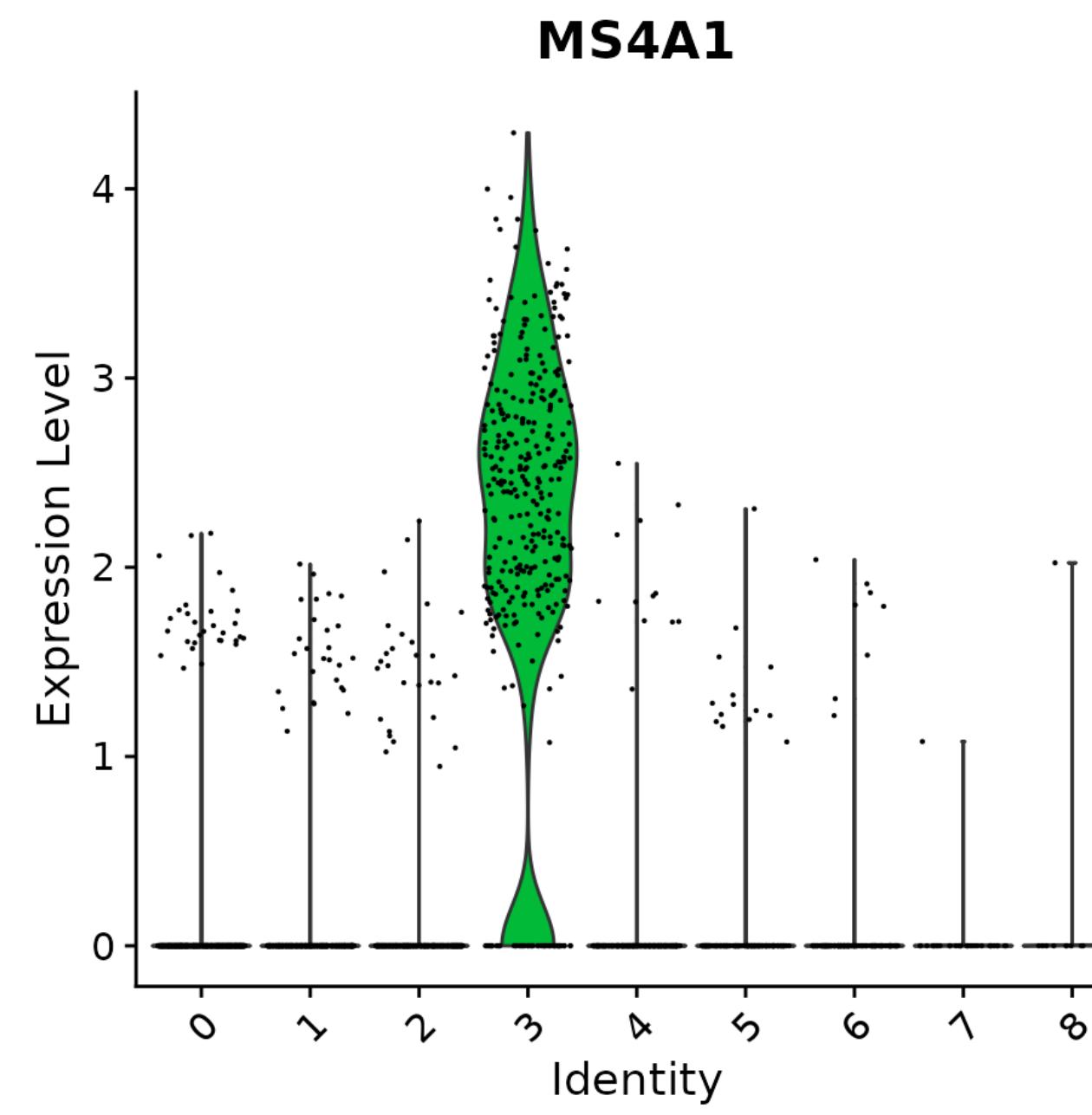
Wie viele Clusters?

Dazu verwenden wir die Darstellung der Hauptkomponenten aus der PCA und deren jeweilige Erklärungskraft. Der Punkt, an dem wir den Knick sehen, ist der Cut-out-Wert. Es ist nicht notwendig, weitere Dimensionen zu berücksichtigen, da diese keinen wesentlichen Beitrag zur Erklärung der Daten mehr leisten.



Downstream Analyse

Marker-Gene finden, die Cluster über differentielle Expression definieren



Downstream Analyse

Zuweisung der Zelltyp-Identität zu Clustern auf Grundlage der Literatur: Erstellung einer Expressions-Heatmap für bestimmte Zellen und Merkmale für die wichtigsten Marker.

