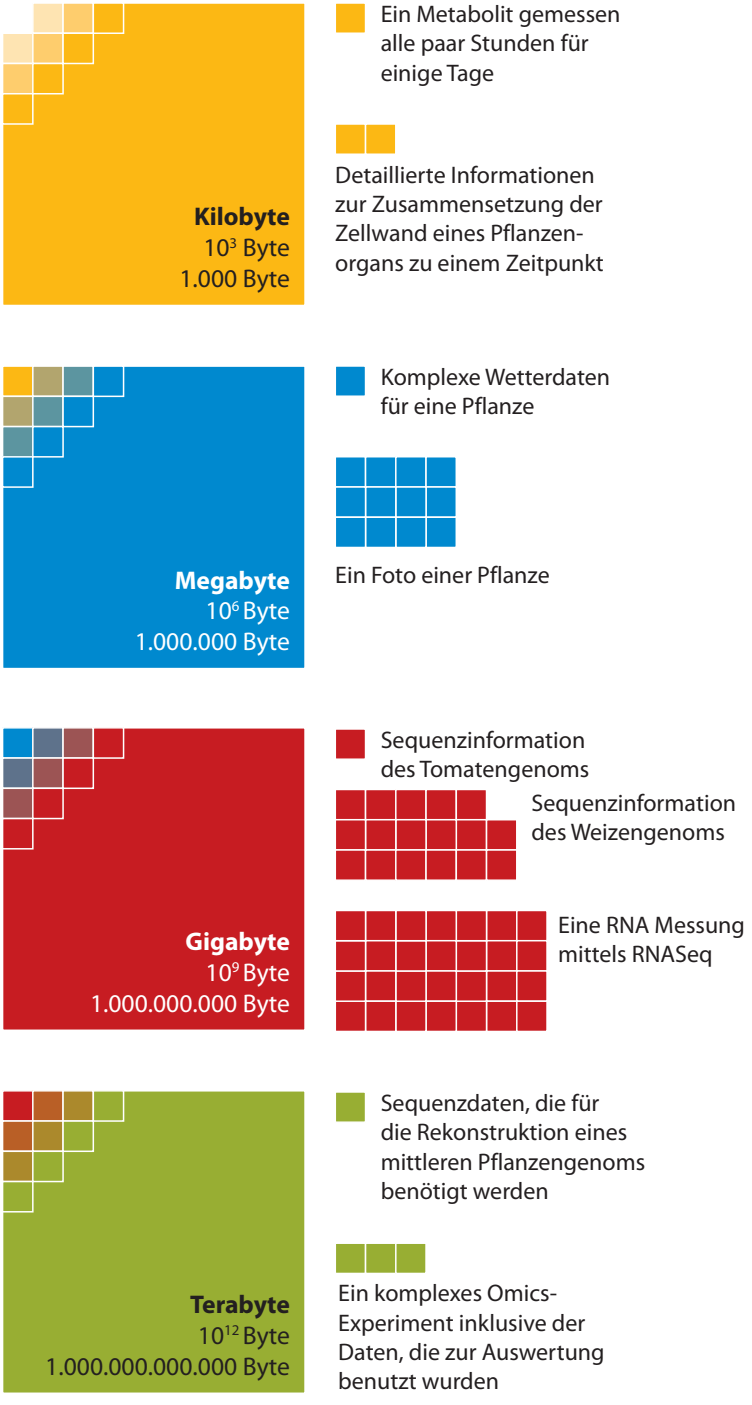
Arbeitsmaterial **Modul 2**

Kilo, Mega, Giga, Tera…

Wenn in der Wissenschaft Experimente im Labor oder in der freien Natur durchgeführt werden, und beispielsweise der Stoffwechsel einer Pflanze untersucht wird, entstehen groß Mengen an Forschungsdaten. Die Größe dieser Datensätze kann von wenigen Kilobyte bis zu vielen Terabyte und mehr reichen. Sie hängt von Faktoren, wie der Anzahl der untersuchten Merkmale, der Länge des Beobachtungszeitraums, den Datenformaten (Bild, Text, Video), aber auch von der Pflanze selbst ab. Wenn ein einzelnes Stoffwechselprodukt (Metabolit) über einen kurzen Zeitraum beobachtet, die Zellwandstruktur zu einem bestimmten Zeitpunkt analysiert oder das Reaktionsverhalten von Pflanzen auf äußere Bedingungen betrachtet wird, werden die erhobenen Messpunkte in Listen bzw. Tabellen festgehalten, die weniger Speicherplatz benötigen als beispielsweise ein hochauflösendes Foto.

Die Entschlüsselung und Analyse des Erbguts ist um einiges umfangreicher, weil die Anzahl der untersuchten Elemente in die Millionen und Milliarden geht. Bei der DNA-Sequenzierung wird die Abfolge der Basen (Adenin, Guanin, Thymin und Cytosin) im Genom entschlüsselt. Je mehr Basenpaare vorhanden sind, desto größer ist auch der Datensatz: Tomaten besitzen etwa 35.000 Gene mit rund 900 Megabasenpaaren, was 900 Millionen Basenpaaren entspricht. Beim hexaploiden Weizen sind es dagegen 94.000 bis 96.000 Gene mit 17 Gigabasenpaaren, umgerechnet also 17 Milliarden Basenpaaren.

Während es bei der DNA-Sequenzierung um die Erfassung der Basensequenz geht, liefert die RNA-Analyse mittels RNA-Seq (Gesamt-Transkriptom-Shotgun-Sequenzierung) Informationen über die Aktivität der Gene (Genexpression) bestimmter Pflanzenbereiche oder unterschiedlicher Entwicklungsstadien. Da jedoch nicht nur ein Interesse an rezenter (heute verfügbarer) DNA besteht, sondern z.B. auch an ausgestorbenen Arten, deren Erbgut nicht oder nur teilweise erhalten ist, spielt die Rekonstruktion von fossilen Genomen eine wichtige Rolle. Um Aspekte der Evolution verstehen zu können, muss das Erbgut von mehreren Arten als auch das von unterschiedlichen Individuen der gleichen Art herangezogen werden. Weiterhin müssen genetische Informationen, die nicht im Zellkern liegen und herrschende Umweltbedingungen einbezogen werden, wodurch sich das Datenvolumen zusätzlich erhöht.



© Björn Usadel / RWTH Aachen, FZ Jülich

Dass Omics-Experimente so umfangreich sind und die größten Datenmengen produzieren liegt daran, dass bei der umfassenden Betrachtung der Gesamtheit der Gene (*Genomics*), des Stoffwechsels (Metabolomics) oder der Proteine (*Proteomics*) nicht mehr nur die Einzelelemente für sich untersucht werden, sondern zusätzlich auch die mit ihnen verbundenen komplexen Prozesse und Mechanismen. Die Daten besitzen oft auch eine zeitliche Dimension und beinhalten Informationen über die Vernetzung und Regulation der einzelnen Elemente. Leistungsstarke Computer, die Bioinformatik und die Systembiologie schaffen die Voraussetzungen für derartige Analysen. Ohne solche Werkzeuge wäre zum Beispiel die Entschlüsselung des menschlichen Genoms im Jahr 2001, das Humangenomprojekt, nicht möglich gewesen.