Sandro Andreotti
Bioinformatics Solution Center

Freie Universität Berlin

Projektmanagement (Softwarepraktikum)

Thema: Workflows

Typisches Szenario in der Praxis

Benötigt: Auswertung biologischer Massendaten z.B.
- NGS
- Massenspektrometrie
- Mikroskopie

(Leider) selten!

Es existiert ein Tool für die gesamte Analyse!

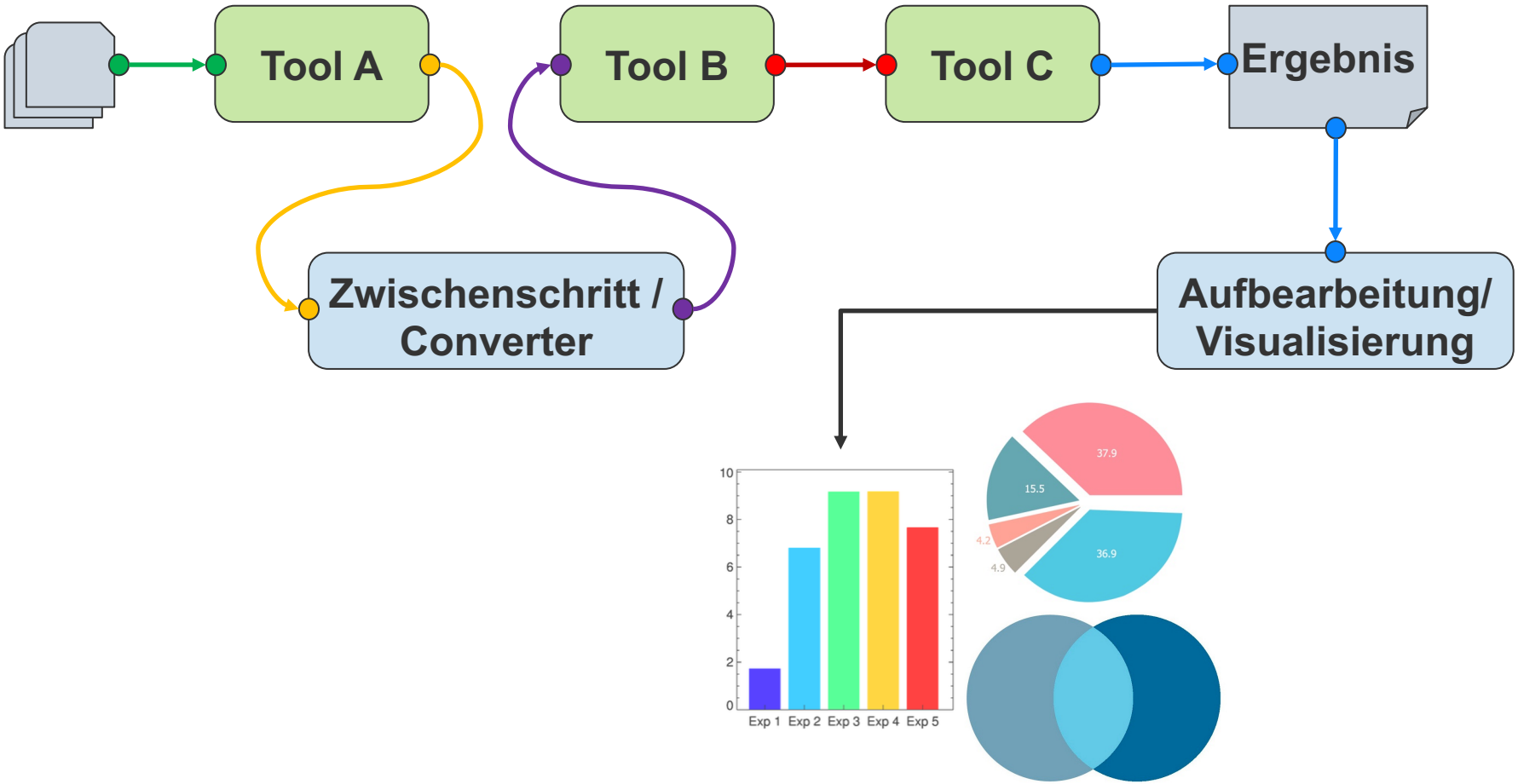
(Zum Glück) selten!

Es existiert noch gar keine Software! Ich muss alles selbst entwickeln!

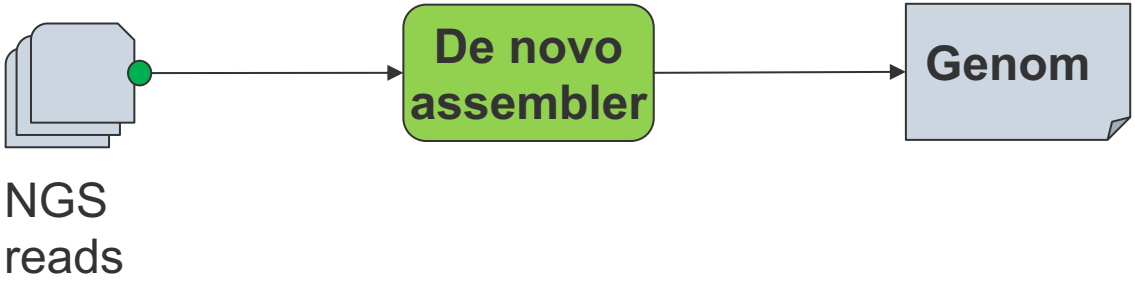
Häufig!

Es existieren Tools für einzelne Schritte der Analyse!

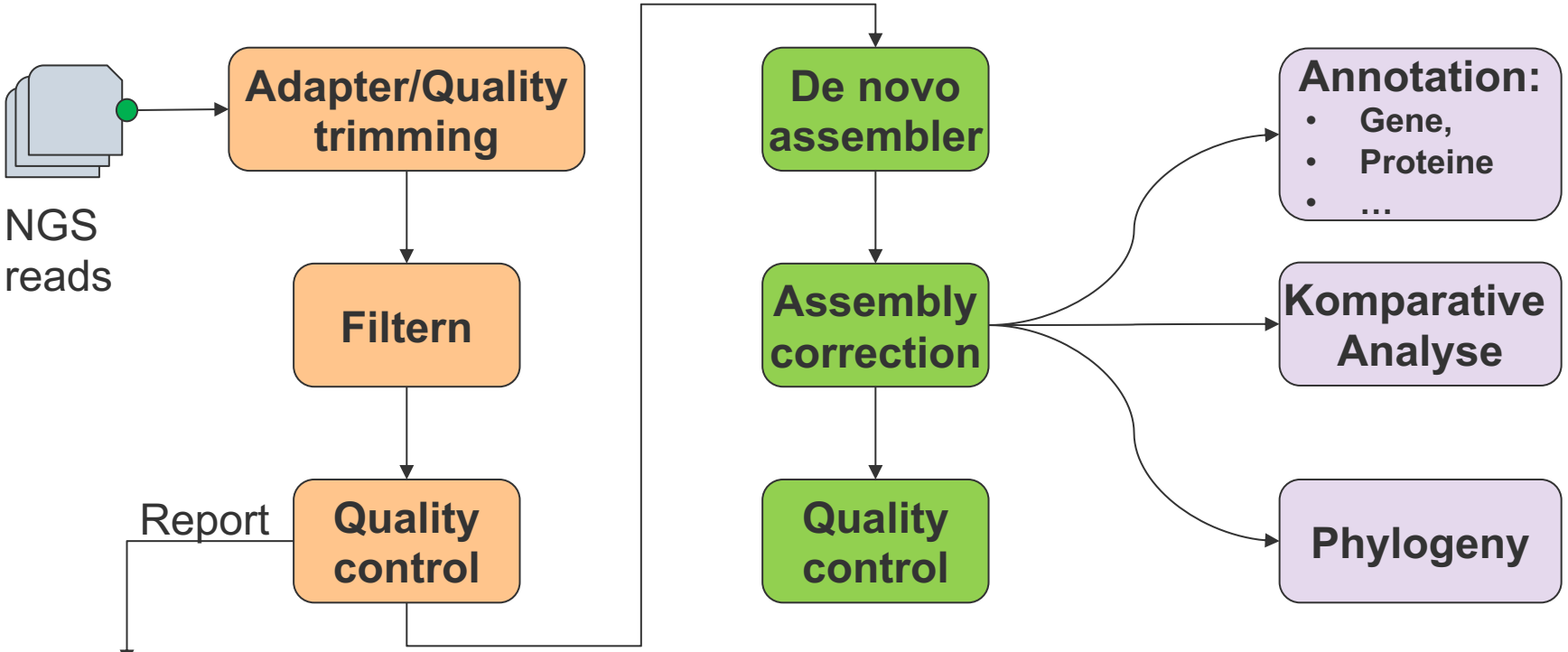
Einzelne Tools → Analyse Pipeline

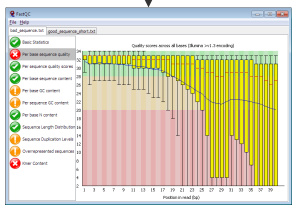


Beispiel: Genom Assemblierung



Beispiel: Genom Assemblierung

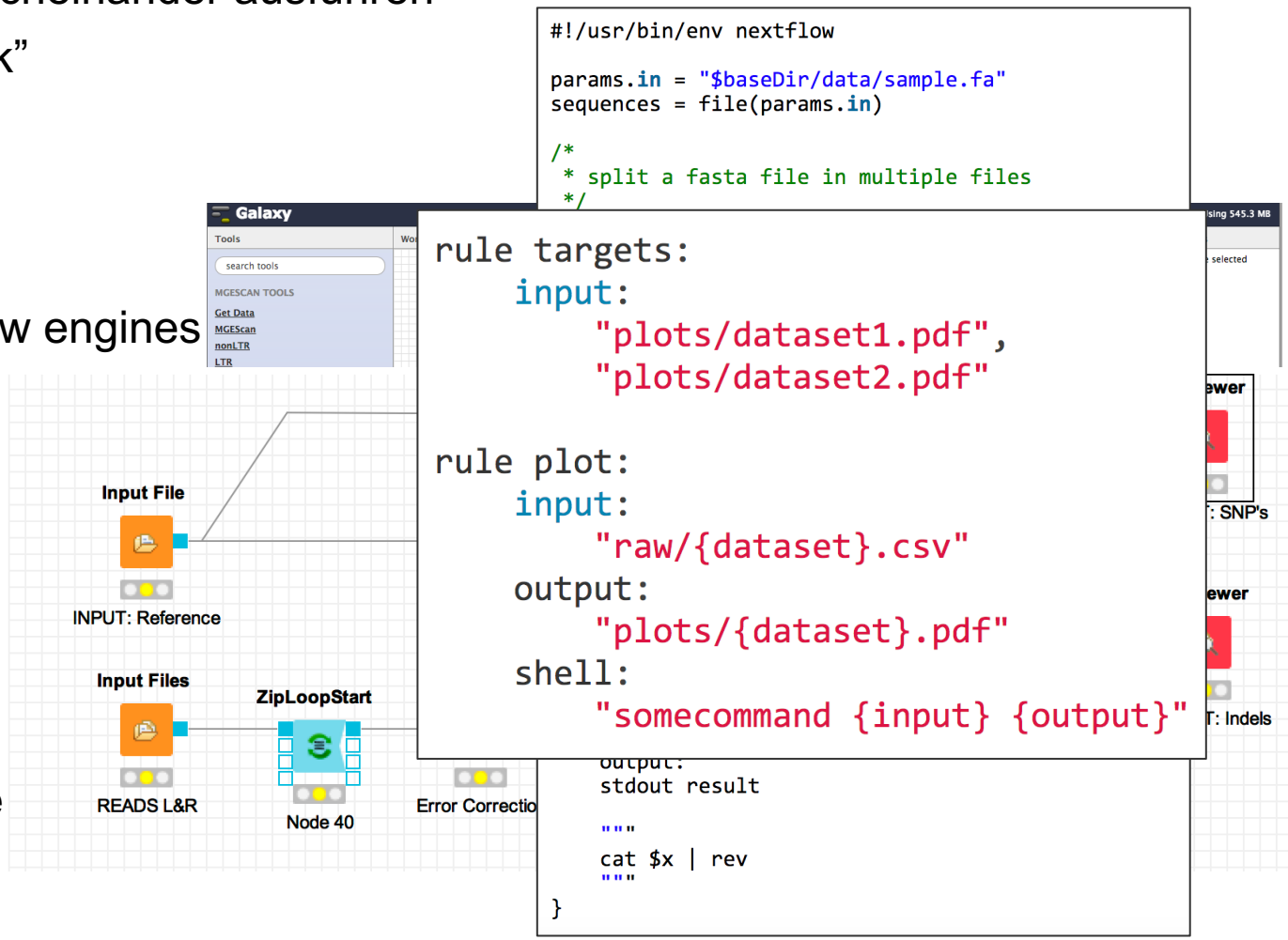




From: FastQC

Realisierung

1. Von Hand Tools nacheinander ausführen
2. Eigenes “framework”
   - Batch script
   - Python script
   - ...
3. Generische workflow engines
   - GUI basiert:
     - Galaxy
     - KNIME
   - Script basiert:
     - Nextflow
     - **Snakemake**



Projekte

Analyse „echter Daten“ aus „echten Projekten“ am BSC

- **Konzeption** einer umfassenden Analyse-Pipeline
  - Recherche:
    - Welche Tools gibt es, was derzeit *best practice?*
    - Wie müssen sie miteinander verbunden werden?
    - Benötigen wir spezielle Zwischenschritte → eigene Programme / Skripte?

- **Implementierung** der Pipeline und ggf. eigener Programme / Skripte
  - Pipeline in **Snakemake**
  - Eigene Programme nach Wahl (Python, R, C++,...)

- **Testen** anhand der Daten + Auswertung und Diskussion mit den Wissenschaftlern
  - **Verfeinerung / Erweiterung** anhand von Feedback.

Projekte

Analyse „echter Daten“ aus „echten Projekten“ am BSC

Beispiel Projekt A:

Multi-Spezies RNA-Seq Analyse:

- **Daten**: RNA-Seq Daten aus verschiedenen Geweben
- **Fragen**: Für welche Gene unterscheidet sich die Expressionen zwischen Geweben.
- **Problem:** Die *„biologischen Replikate“* sind verschiedene aber nah verwandte Spezies.
- **Aufgabe**: Anpassung etablierter RNA-Seq Analyseworkflows an die multi-Spezies Situation.

Projekte

Analyse „echter Daten“ aus „echten Projekten“ am BSC

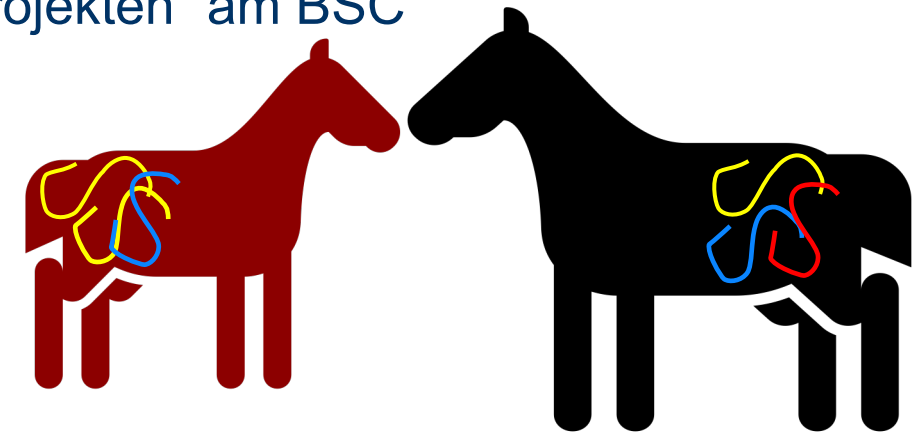
Beispiel Projekt A:

Multi-Spezies RNA-Seq Analyse:

- **Ideen:**
  - Mapping-basiert:
    - Mapping auf gemeinsames Referenzgenom
    - Polishing des Genoms für jede Spezies → angepasste Referenz pro Spezies
    - Finales Mapping und Check ob angepasste Referenz gut genug ist.
  - Assembly-basiert:
    - De-novo Transkriptom Assembly pro Spezies.
    - Matching der Transkripte aus verschiedenen Spezies
    - Annotation der gematchten Transkripte

Projekte

Analyse „echter Daten“ aus „echten Projekten“ am BSC



Beispiel Projekt B:

Wurmpopulation bei Pferden:

- Pferde an unterschiedlichen Standorten
- Fragen: Wie unterscheidet sich die Spezies-Zusammensetzung der Cyathostominae zwischen verschiedenen Standorten

- **Aufgabe:**
  - Anhand von **amplicon** *NGS-Daten* des mitochondrialen COX1 Gens soll die Zusammensetzung der Population für jeden Standort ermittelt werden.
  - Suche nach signifikanten Unterschieden zwischen verschiedenen Standorten (sowie weiteren Merkmalen wie z.B. Behandlung)

Projekte

Analyse „echter Daten“ aus „echten Projekten“ am BSC

Beispiel Projekt C:

Optimierung QuantSeq Pipeline:

- QuantSeq ist RNA-Seq am 3‘ Ende
- Aufgabe: Optimierung des RNA-Seq-Workflows für Lexogen QuantSeq Daten.

- **Aufgaben:**
  - Vorverarbeitung des Referenzgenoms durch z.B. Extraktion der 3‘ Bereiche.
  - Implementierung und Vergleich verschiedener DGE Ansätze (DESeq2, Kallisto-Sleuth, Salmon, edgeR.
  - ...

Projekte

Analyse „echter Daten“ aus „echten Projekten“ am BSC

Viele mögliche weitere Projekte...

- Metagenomanalyse
- Genom Assemblierung
- (Funktionelle) Genomannotation

Eckdaten

Zeitplan (FLEXIBEL)	
Vorbesprechung	Mitte Februar – Anfang März
Seminar (Snakemake, Git, NGS-Tools,...)	Mitte/Ende März
Vorstellung der Projektpläne	Anfang April
Eigenständige Bearbeitung und wöchentliche Besprechung	April + Mai
Vorstellung und Abschlussbericht	Ende Mai

Quantitative Aufteilung: (in %)
Praktische Programmierarbeit: 50%
Soft Skills: 50%
Verwendete Programmiersprache(n):
R, Python oder andere Skriptsprache

Schwierigkeitsgrad
A Programmieren ★★★★
B Biologie/Chemie ★
C Projektmanagement ★★★

Erforderliche Vorkenntnisse:
R, **Python**