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!

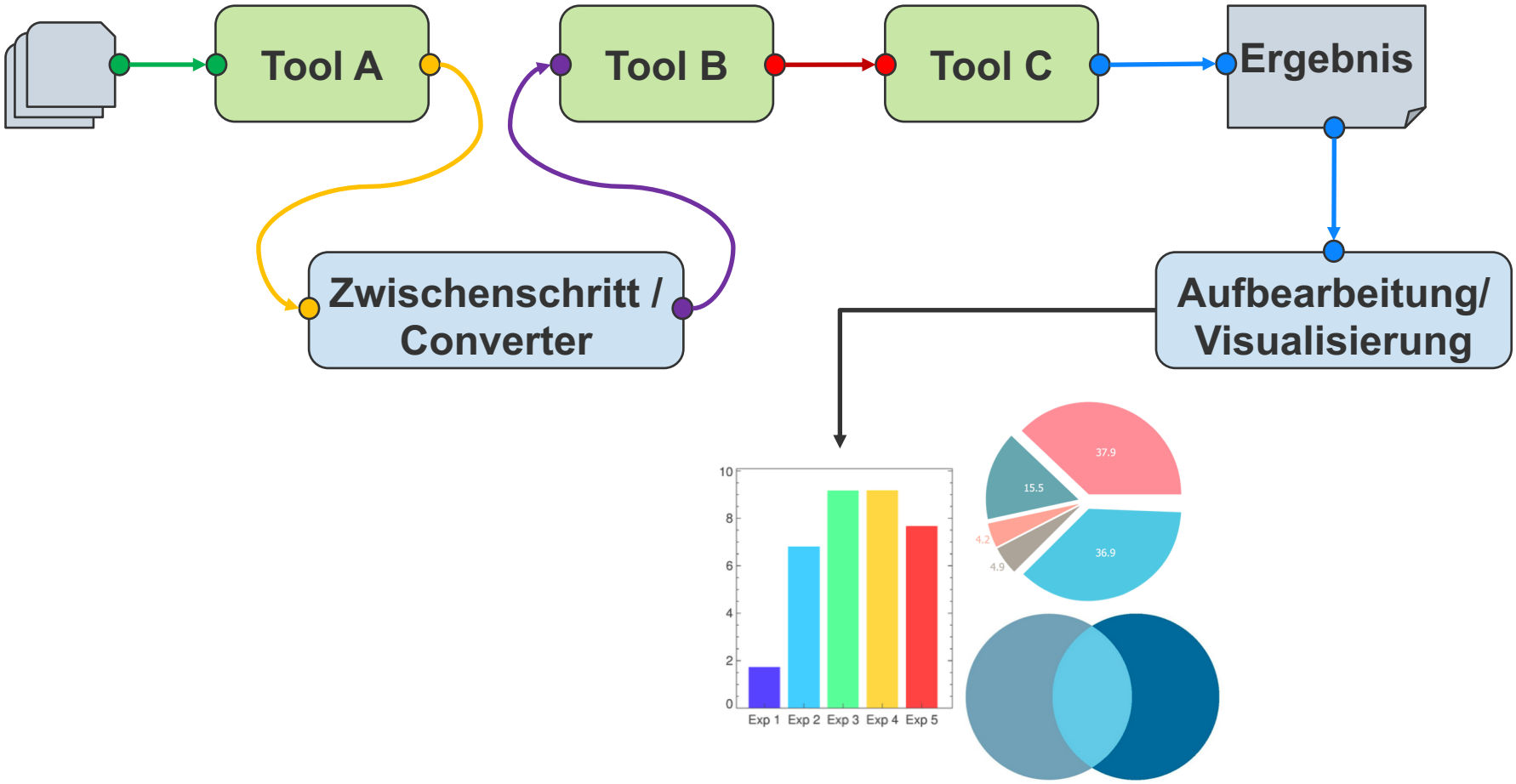
(Zum Glück) selten!

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

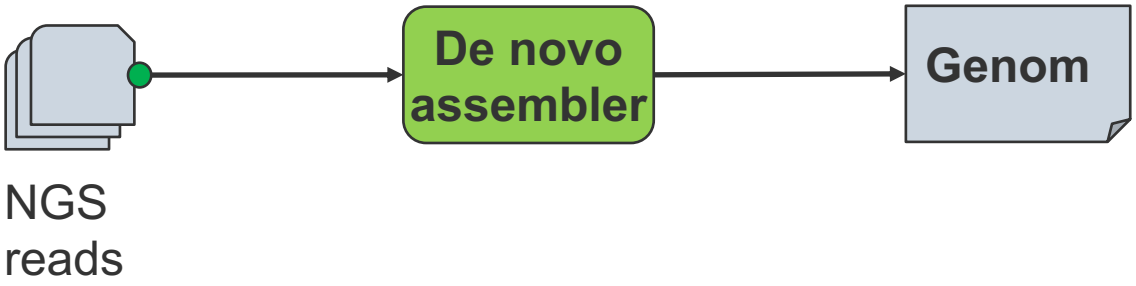
Häufig!

Es existieren Tools für einzelne Schritte der Analyse!

Einzelne Tools → Analyse Pipeline

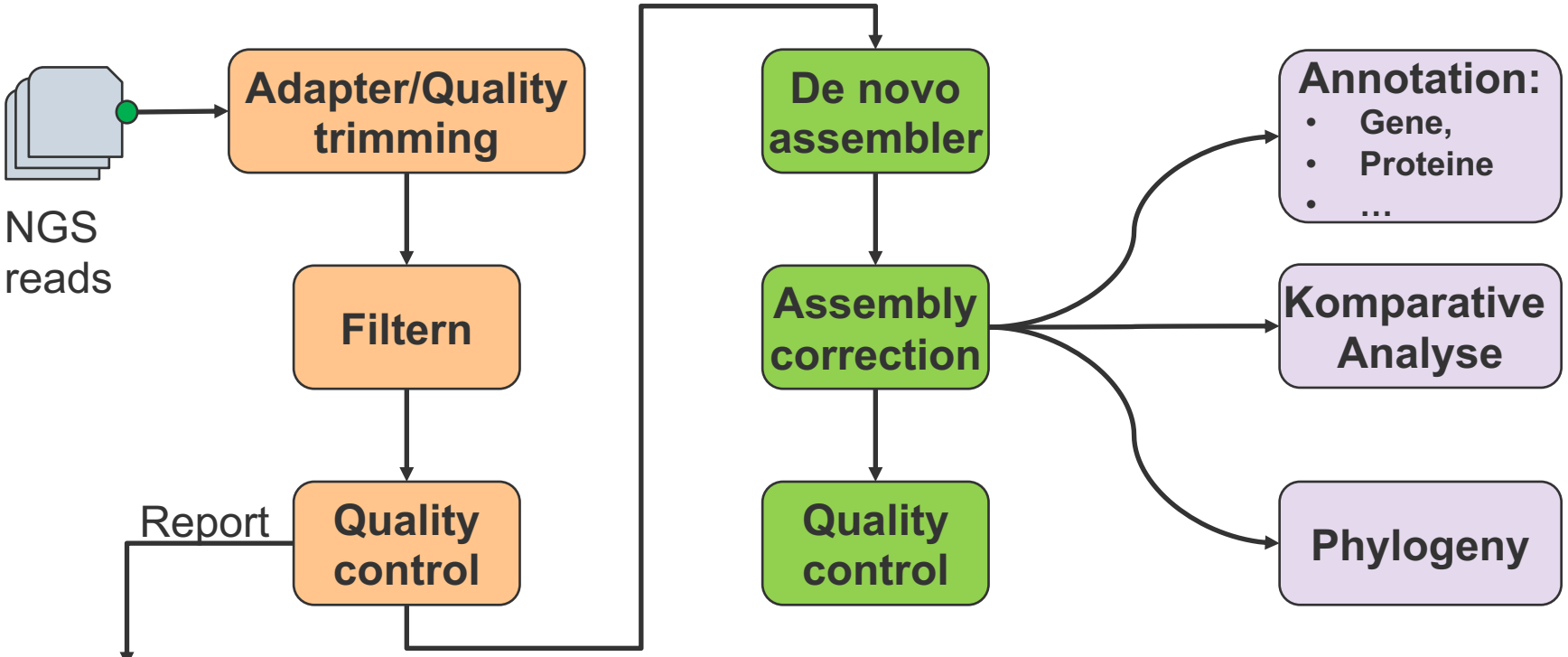


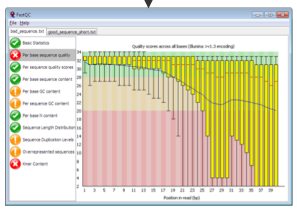
Beispiel: Genom Assemblierung



?

Beispiel: Genom Assemblierung





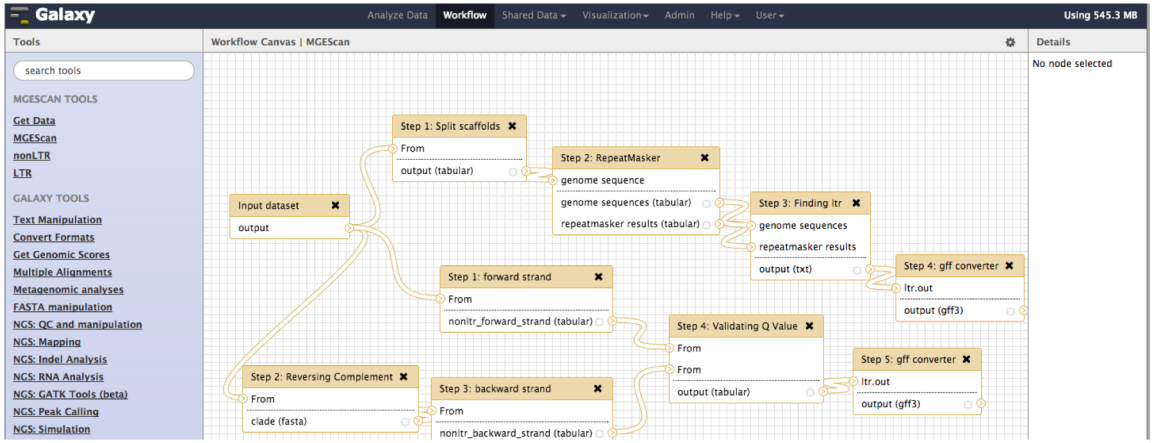
From: FastQC

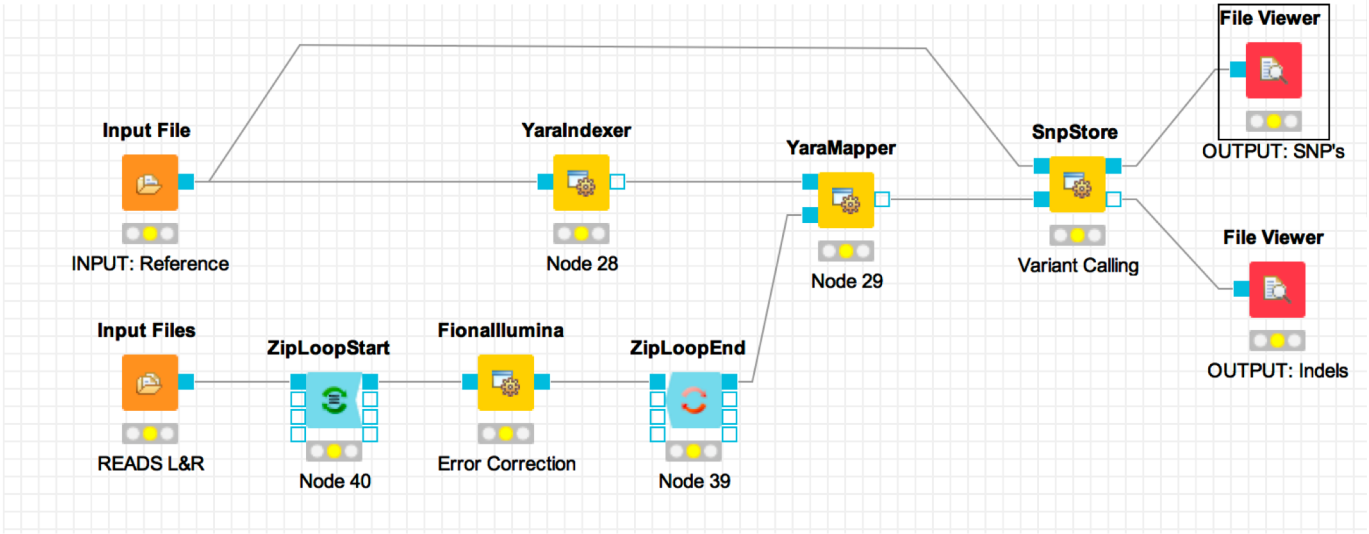
Realisierung

1. Von Hand Tools nacheinander ausführen
2. Eigenes “framework”
   - Batch script
   - Python script
   - ...

Realisierung

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





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**

```
#!/usr/bin/env nextflow

params.in = "$baseDir/data/sample.fa"
sequences = file(params.in)

/*
 * split a fasta file in multiple files
 */
process splitSequences {

    input:
    file 'input.fa' from sequences

    output:
    file 'seq_*' into records

nt > f}' < input.fa
```

```
rule targets:
    input:
        "plots/dataset1.pdf",
        "plots/dataset2.pdf"

rule plot:
    input:
        "raw/{dataset}.csv"
    output:
        "plots/{dataset}.pdf"
    shell:
        "somecommand {input} {output}"
```

Projekte

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

- **Konzeption** einer umfassenden Analyse-Pipeline
  - Recherche:
    - Welche Tools gibt es, was ist 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, C++,...)

- **Testen** anhand der Daten + Auswertung und Diskussion der Ergebnisse mit den Wissenschaftlern („Kunden“)
  - **Verfeinerung / Erweiterung** anhand von Feedback.

Projekte

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

Beispiel Projekt A:

Wurm-Biom Analyse von Pferden:

- Wurminfektionen sehr häufig bei Pferden
- Ein Wirt kann gleichzeitig von vielen Spezies betroffen sein.

- **Aufgabe:**
  - Anhand von *NGS-Daten* von Wurm-DNA aus Pferden verschiedener Herkunft und verschiedener Behandlung die Spezies-Zusammensetzung bestimmen und quantifizieren.

Projekte

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

Beispiel Projekt B:

Bestimmung von Resistenzgenen mittels RNA-Seq:

- Wieder Pferdewürmer, diesmal multi-resistente Exemplare

- **Aufgabe:**
  - Anhand von *RNA-Seq* Daten für Würmer mit verschiedenen Treatments sollen Gene identifiziert werden, welche als Reaktion auf den Wirkstoff hochreguliert sind → Kandidaten für Resistenzgene.

Projekte

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

Diverse weitere Projekte sind möglich und werden in der

Seminarwoche vorgestellt

SWP - Orga

Zeitplan (vorläufig, kann angepasst werden)	
Vorbesprechung	Ende Februar
Seminar (Snakemake, Git, NGS-Tools,...)	9.3. – 13.3.
Vorstellung der Projektpläne	25.3
Eigenständige Bearbeitung und wöchentliche Besprechung	Freitags (10-12) und nach Bedarf
Bearbeitung und Abschlussbericht	bis 8.5.

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