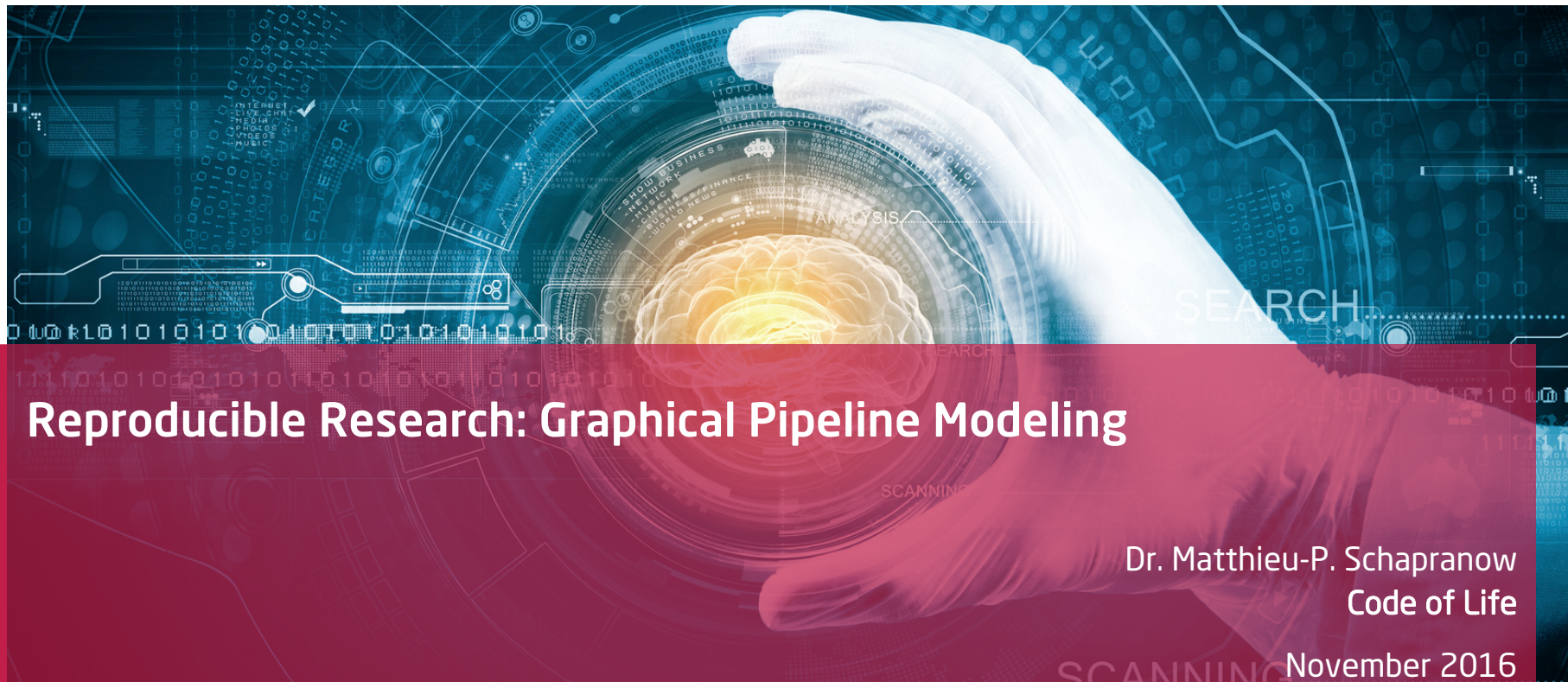




Reproducible Research: Graphical Pipeline Modeling

Dr. Matthieu-P. Schapranow
Code of Life

November 2016

Genome Data Processing Pipelines

State of the Art



```
bwa aln hg19.fa sample.fastq | bwa samse hg19.fa - sample.fastq | samtools view -Su - | samtools sort ...
```

- Concatenation of command line tools reading/writing files
- Lack of standardization and exchangeability
- Requires dedicated expertise for
 - Setup and configuration,
 - Error handling, and
 - Scalable processing
- Requirement: Enable modeling, configuration, and execution of reproducible pipelines without involving IT experts

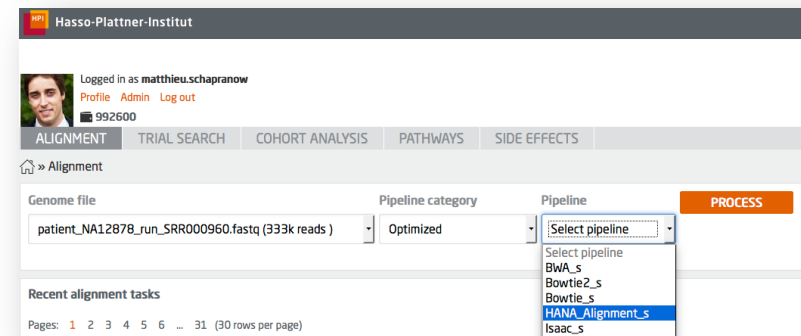
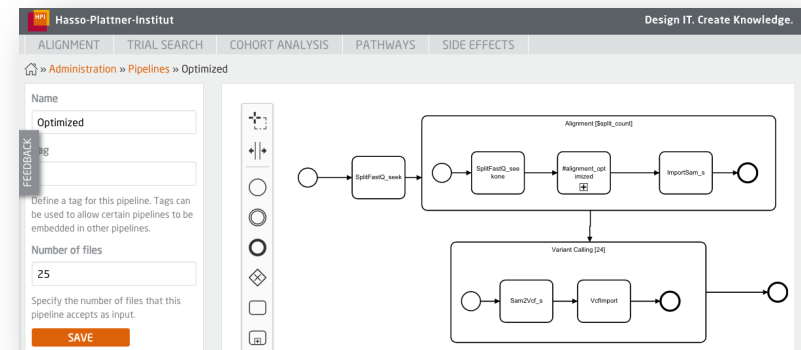
**Reproducible
Research: Graphical
Pipeline Modeling**

Schapranow, Nov 2016

Our Approach From Model to Execution

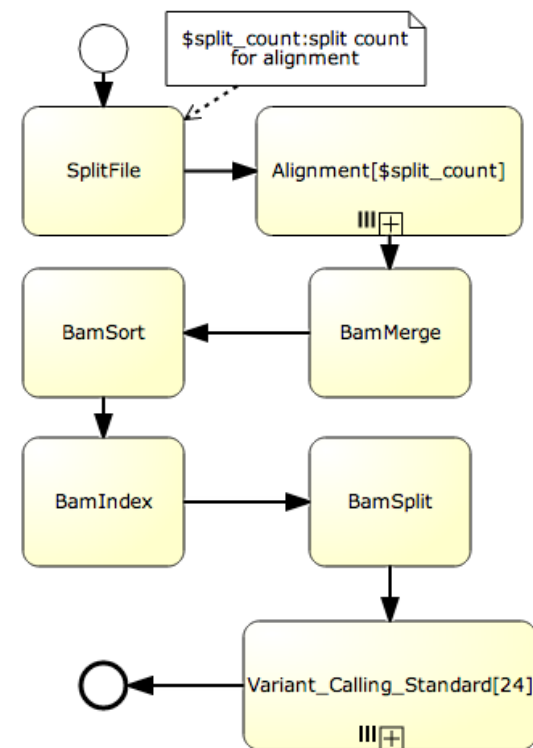


1. Design time (researcher, process expert)
 - Definition of parameterized process model
 - Uses graphical editor and jobs from repository
2. Configuration time (researcher, lab assistant)
 - Select model and specify parameters, e.g. aln opts
 - Results in model instance stored in repository
3. Execution time (researcher)
 - Select model instance
 - Specify execution parameters, e.g. input files



Standardized Graphical Modeling

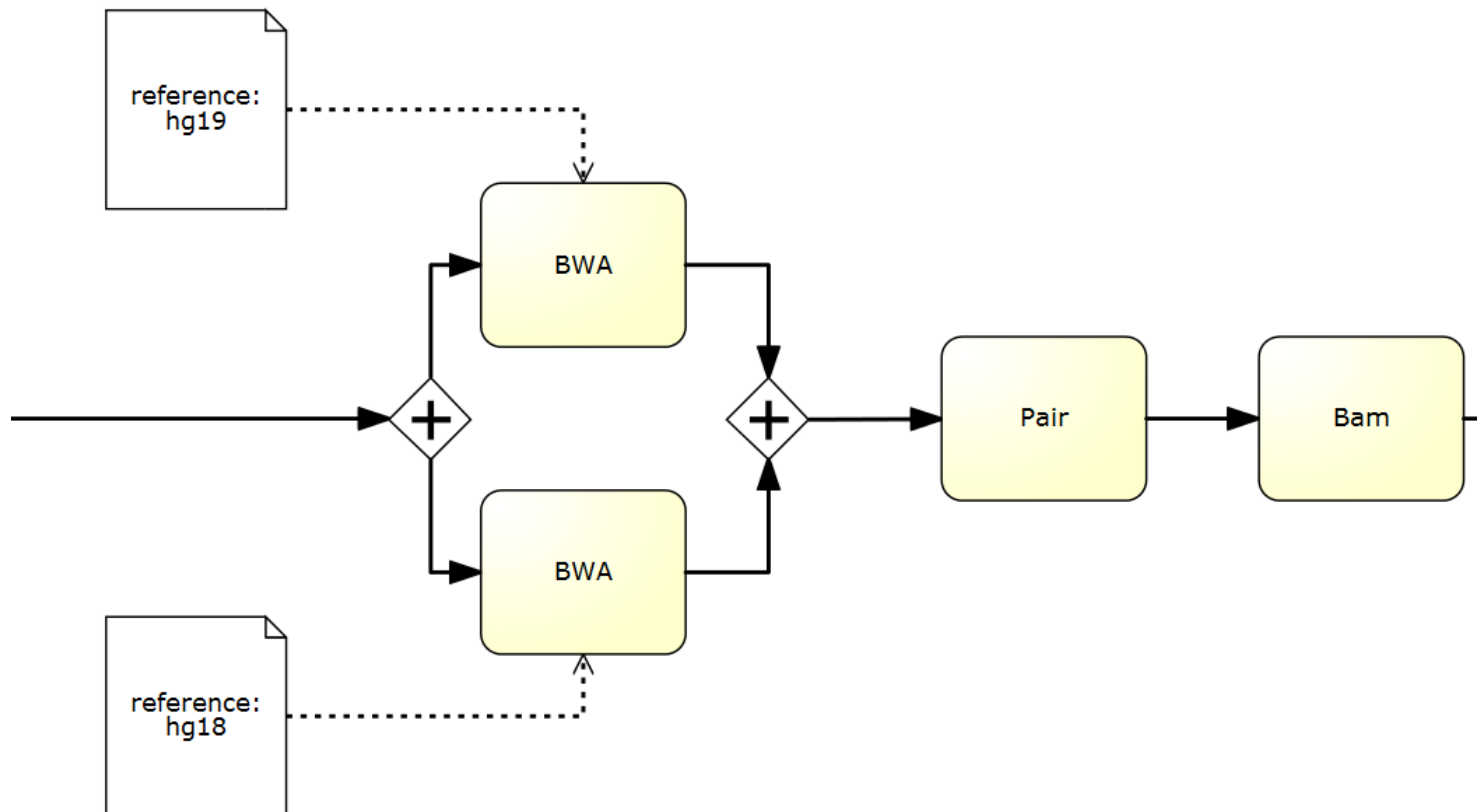
- Graphical modeling notation compliant with Business Process Modeling and Notation 2.0 extended by
 - Modular structure
 - Parallelization annotations
 - Parameters and variables
- Model descriptions are stored within IMDB
- Model instances are transformed into graph structure for execution by dedicated runtime environment



**Reproducible
Research: Graphical
Pipeline Modeling**

Schapranow, Nov 2016

BPMN Example



**Reproducible
Research: Graphical
Pipeline Modeling**

Schapranow, Nov 2016

Persisting Pipelines

XML Process Definition Language

```
<xpdl:Activity CompletionQuantity="1" Id="newpkg1_wp1_act2" Name="BWA">
  <xpdl:Implementation>
    <xpdl:No/>
  </xpdl:Implementation>
  <xpdl:Performers>
    <xpdl:Performer>newpkg1_wp1_par1</xpdl:Performer>
  </xpdl:Performers>
  <xpdl:NodeGraphicsInfos>
    <xpdl:NodeGraphicsInfo BorderColor="#000000" FillColor="#99FF99">
      <xpdl:Coordinates XCoordinate="239.0" YCoordinate="219.0"/>
    </xpdl:NodeGraphicsInfo>
  </xpdl:NodeGraphicsInfos>
</xpdl:Activity>
```

```
<xpdl:Artifacts>
  <xpdl:Artifact ArtifactType="DataObject" Id="newpkg1_1" Name="newpkg1_1">
    <xpdl:DataObject Id="newpkg1_1" Name="reference:hg19"/>
    <xpdl:NodeGraphicsInfos>
      <xpdl:NodeGraphicsInfo BorderColor="#000000" FillColor="#E8EEF7">
        <xpdl:Coordinates XCoordinate="239.0" YCoordinate="74.0"/>
      </xpdl:NodeGraphicsInfo>
    </xpdl:NodeGraphicsInfos>
  </xpdl:Artifact>
```

Database Structure

PIPELINES.MODELS

ID	NAME	CONTENT	TYPE
154	BWA2	<?xml version="1.0" encoding="UTF-8"?>	alignment
150	VCFImport	<?xml version="1.0" encoding="UTF-8"?>	
146	Alignme...	<?xml version="1.0" encoding="UTF-8"?>	
144	Bowtie2_s	<?xml version="1.0" encoding="UTF-8"?>	alignme...
143	TMAP4	<?xml version="1.0" encoding="UTF-8"?>	alignment
141	TMAP2	<?xml version="1.0" encoding="UTF-8"?>	alignment
140	TMAP1	<?xml version="1.0" encoding="UTF-8"?>	alignment
139	HANA_Al...	<?xml version="1.0" encoding="UTF-8"?>	alignment
138	BWA	<?xml version="1.0" encoding="UTF-8"?>	alignment
137	Bowtie2	<?xml version="1.0" encoding="UTF-8"?>	alignment
136	Bowtie	<?xml version="1.0" encoding="UTF-8"?>	alignment
135	HANA_Al...	<?xml version="1.0" encoding="UTF-8"?>	alignme...
134	BWA_s	<?xml version="1.0" encoding="UTF-8"?>	alignme...
133	Bowtie_s	<?xml version="1.0" encoding="UTF-8"?>	alignme...
129	Optimized	<?xml version="1.0" encoding="UTF-8"?>	main
128	Standard	<?xml version="1.0" encoding="UTF-8"?>	main
102	Paired_A...	<?xml version="1.0" encoding="UTF-8"?>	main

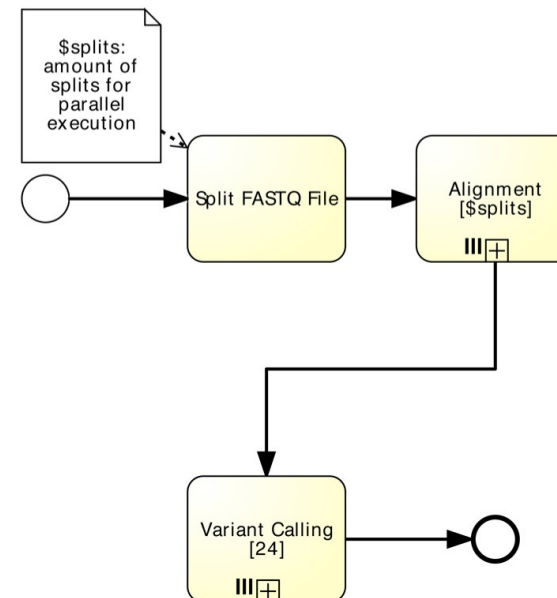
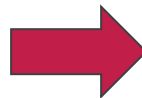
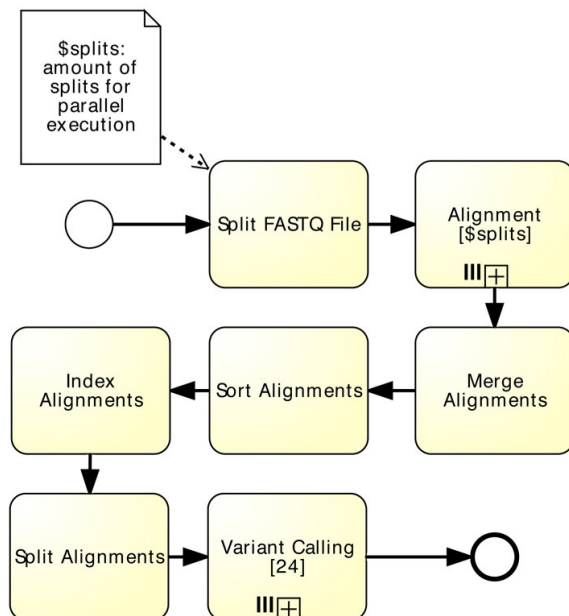
PIPELINES.PIPELINES

ID	NAME	MODE	CONFIG	SUBTASKS
106	Test	129	{"split_count":"25","reference":"hg19"}	{"alignment_speed":"135"}
105	Test5	156	{"reference":"hg19"}	{}
104	Stand...	128	{"split_count":"3","reference":"hg19"}	{"alignment":"154"}
103	test	153	{"reference":""}	{}
78	TMAP4	128	{"split_count":"3","reference":"hg19"}	{"alignment":"143"}
77	TMAP3	128	{"split_count":"3","reference":"hg19"}	{"alignment":"142"}
76	TMAP2	128	{"split_count":"3","reference":"hg19"}	{"alignment":"141"}
74	HANA...	128	{"split_count":"3","reference":"hg19"}	{"alignment":"139"}
73	BWA	128	{"split_count":"3","reference":"hg19"}	{"alignment":"138"}
72	Bowtie2	128	{"split_count":"3","reference":"hg19"}	{"alignment":"137"}
71	Bowtie	128	{"split_count":"3","reference":"hg19"}	{"alignment":"136"}
70	Bowtie...	129	{"split_count":"3","reference":"hg19"}	{"alignment_speed":"144"}
69	HANA...	129	{"split_count":"3","reference":"hg19"}	{"alignment_speed":"135"}
68	BWA_s	129	{"split_count":"25","reference":"hg19"}	{"alignment_speed":"134"}
67	Bowtie_s	129	{"split_count":"3","reference":"hg19"}	{"alignment_speed":"133"}

Schapranow, Nov 2016

Traditional vs. IMDB Approach

- Processing and results are kept within IMDB
- Optimization reduced execution time by >50%



**Reproducible
Research: Graphical
Pipeline Modeling**
Schapranow, Nov 2016

Available Tools

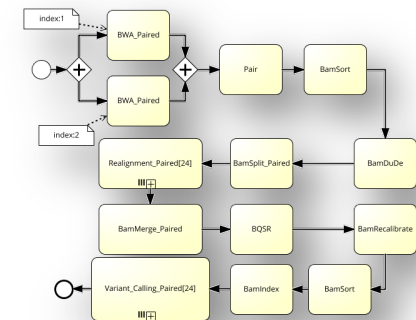
- Pipeline categories (each traditional and optimized):

- Single read,
- Paired read,
- Amplicon

- Alignment: IMDB-based, TMAP, BWA, Bowtie, Isaac

- Variant Calling: IMDB-based, BcfTools, GATK

- Intermediate steps: SAMtools, Picard, GATK

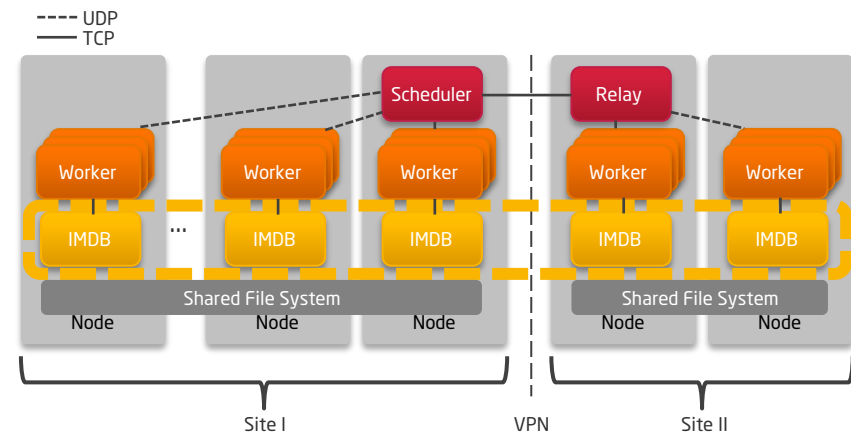


**Reproducible
Research: Graphical
Pipeline Modeling**

Schapranow, Nov 2016

What is Next?

- How to ensure scalable execution of pipelines?



**Reproducible
Research: Graphical
Pipeline Modeling**

Schapranow, Nov 2016

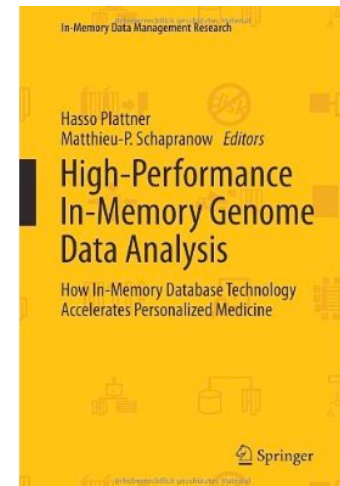
Keep in contact with us!



Dr. Matthieu-P. Schapranow
Program Manager E-Health & Life Sciences

Hasso Plattner Institute
August-Bebel-Str. 88
14482 Potsdam, Germany

schapranow@hpi.de
<http://we.analyzegenomes.com/>



**Reproducible
Research: Graphical
Pipeline Modeling**

Schapranow, Nov 2016