Exome sequencing at the CCG

“NGS an der Schnittstelle von Grundlagen- und translationaler Forschung"

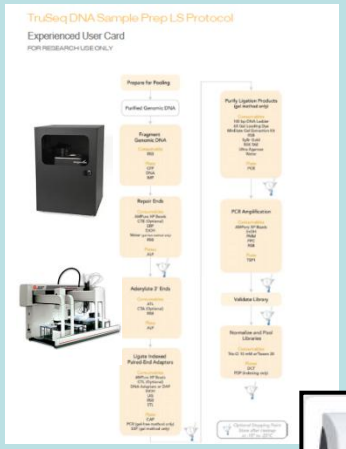
Peter Nürnberg

TMF-Workshop

Berlin, 7.12.2012

The exome sequencing pipeline

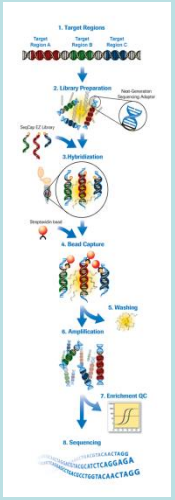
library prep



QC

80samples

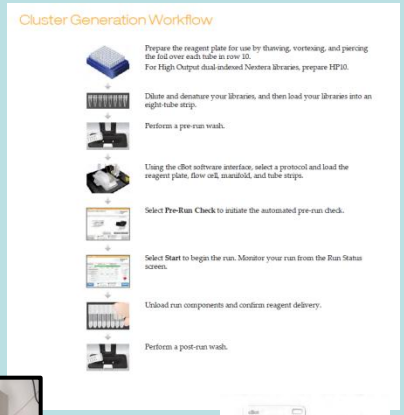
enrichment



qPCR

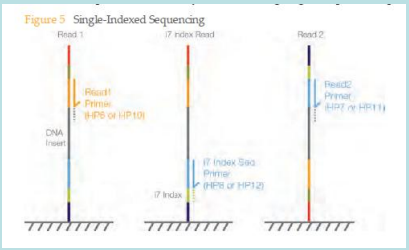
16pools

cluster generation



16 lanes
on 2 flowcells

sequencing



600Gb = raw data
7.5Gb per sample

analysis

DNA QC

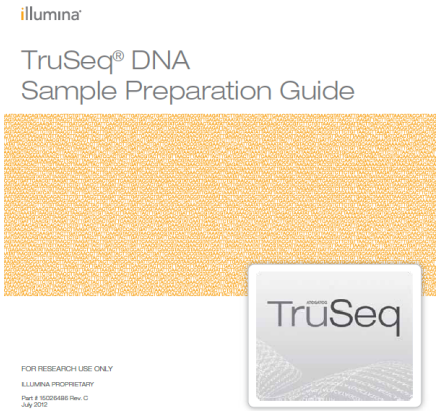
Cologne Center for Genomics

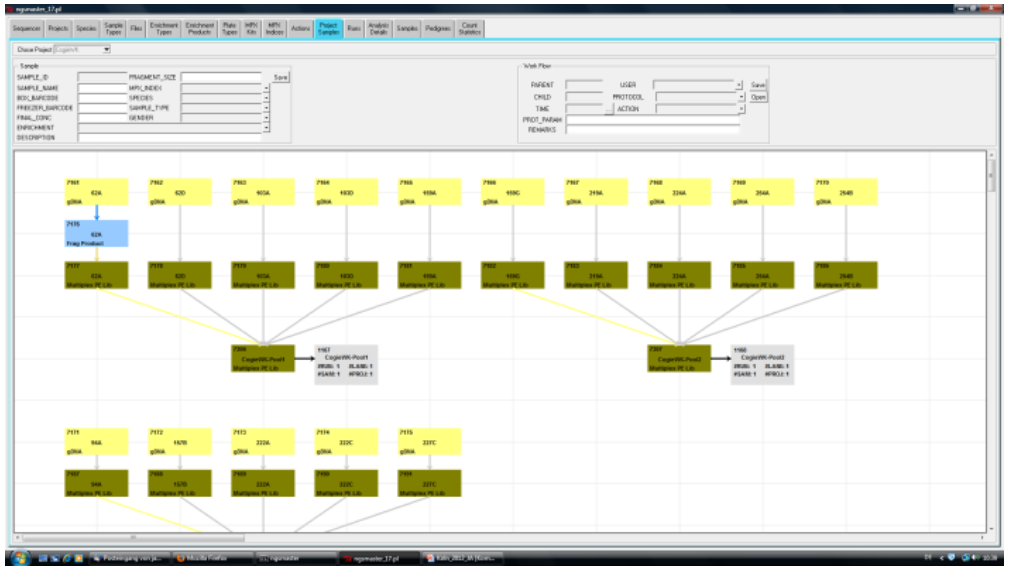
From DNA to library

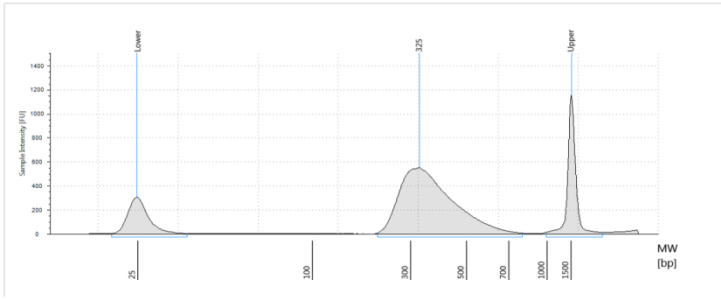
- DNA QC and input in the data base
- Library prep

(fragmentation, end repair , adenylate 3`ends,
Adapter ligation, purification, amplification, validation)

Beckman Coulter. Biomek® FXP







G01_Lib

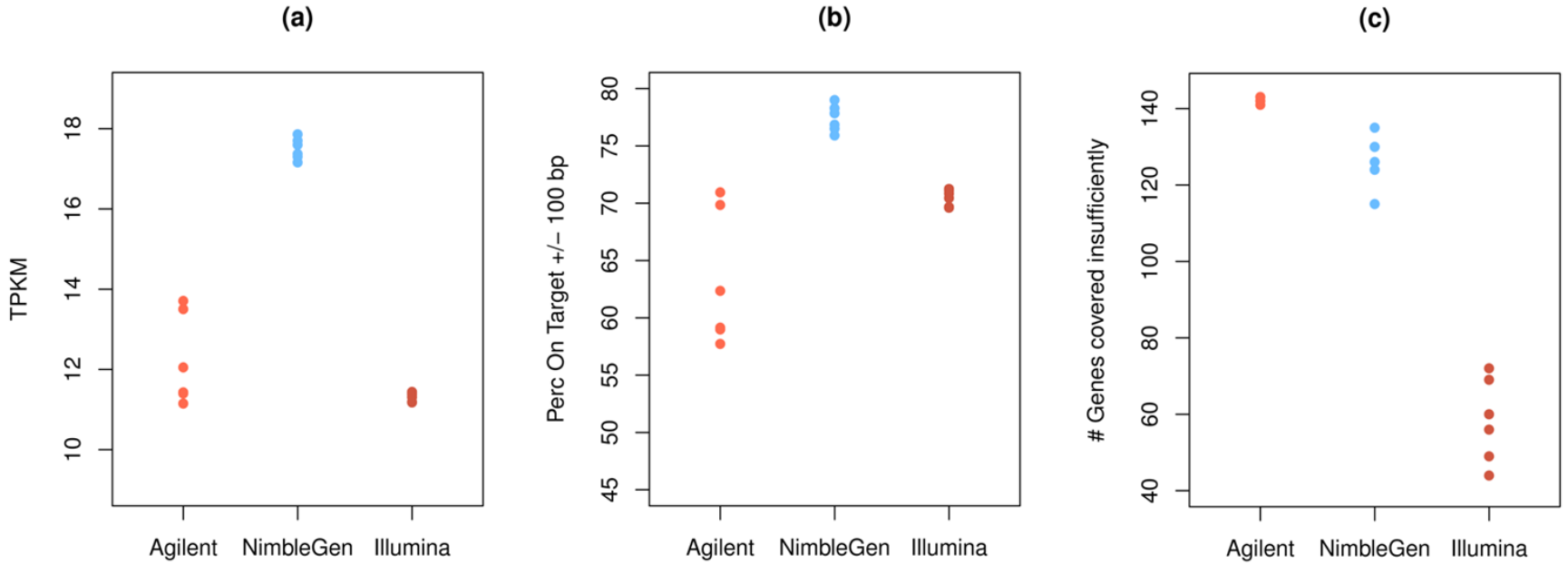
Wavelength	MW [bp]	Conc. [ng/µl]	Molarity [nmol/l]	% of Integrated Area	Peak Comment	Observations
Sample	25	4.46	271			Lower Marker
Sample	325	20.8	96.9	100.00		
Sample	1,500	6.50	6.57			Upper Marker

Enrichment: different products

company	products	Size and content	Ease of use	Performance Q1/2011	Price Q1/2011
Agilent	V1 50Mb V4 V4+UTR	38Mb 51Mb 51Mb 71Mb	x x(+) x x	(3.)	
Nimblegen	Array V2 V3	34Mb 44Mb 64Mb	x x x	(1.)	(+)
Illumina	TruSeq Nextera	62Mb	x	(2.)	

Assessing the enrichment performance in targeted resequencing experiments

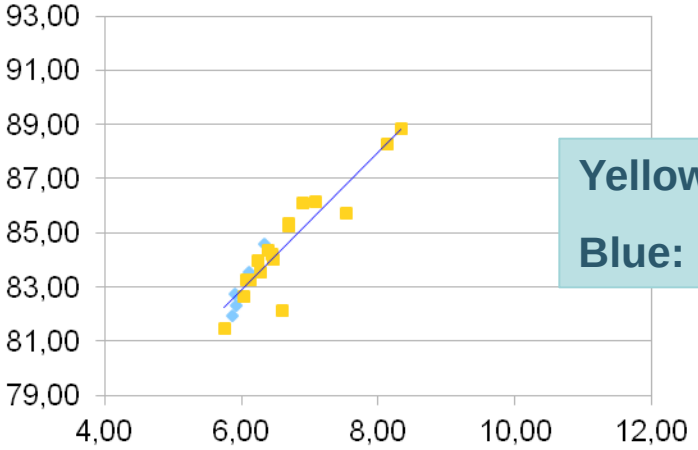
Peter Frommolt et al., Hum Mut 2011



TPKM = On-**T**arget reads **P**er **K**ilobase target region and **M**illion mappable reads

Nimblegen v2: strategy considerations

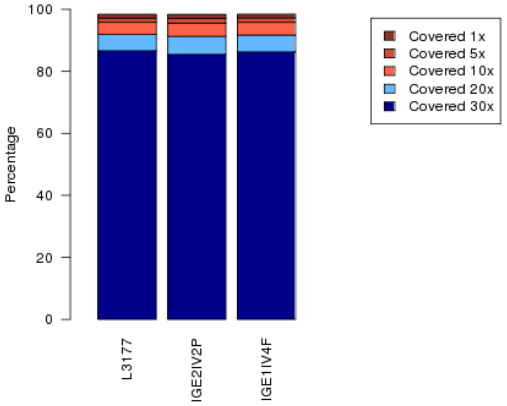
% of target > 30X



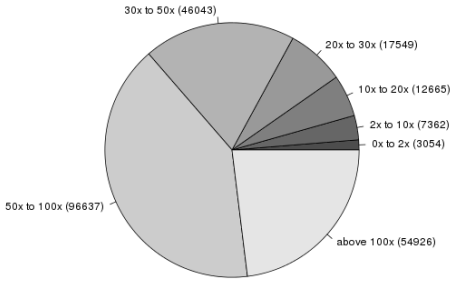
Yellow: pre-capture pooling

Blue: post-capture pooling

[Gb]

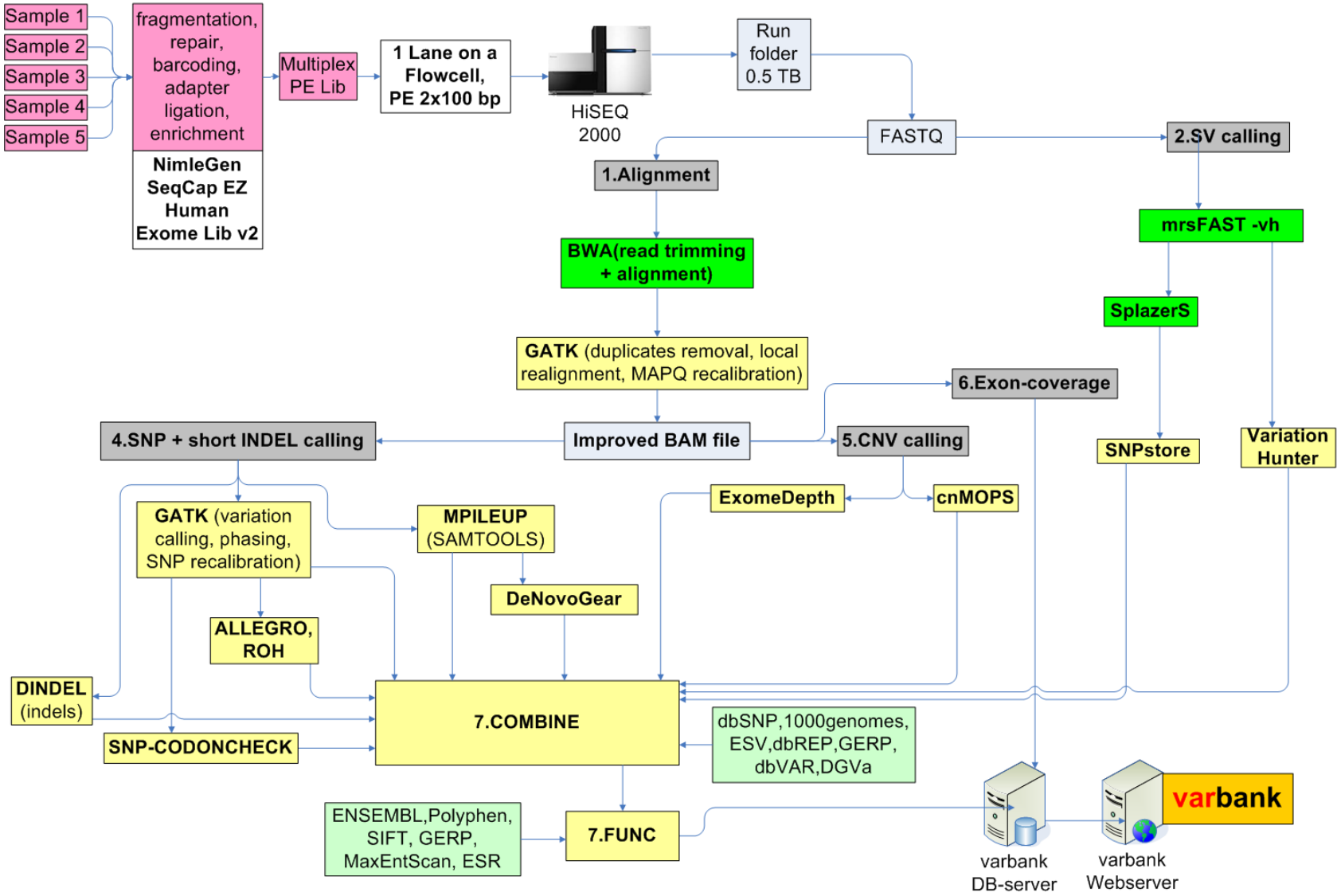


	ideal	real
Raw data	7.5Gb	5-10Gb
aligning	(all)	90-92%
Aligning to target sequence	(all)	72-78%



# Reads	64620968
# On Target ± 100 bp	44549948
Target Size (bp)	44234111
# Target Regions	238236
Coverage Mean	77.27
Coverage Std Dev	64.9
Covered 1x	97.93%
Covered 5x	96.46%
Covered 10x	94.61%
Covered 20x	89.55%
Covered 30x	82.45%
TPKM	15.59

Exome Sequencing Pipeline at the Cologne Center for Genomics

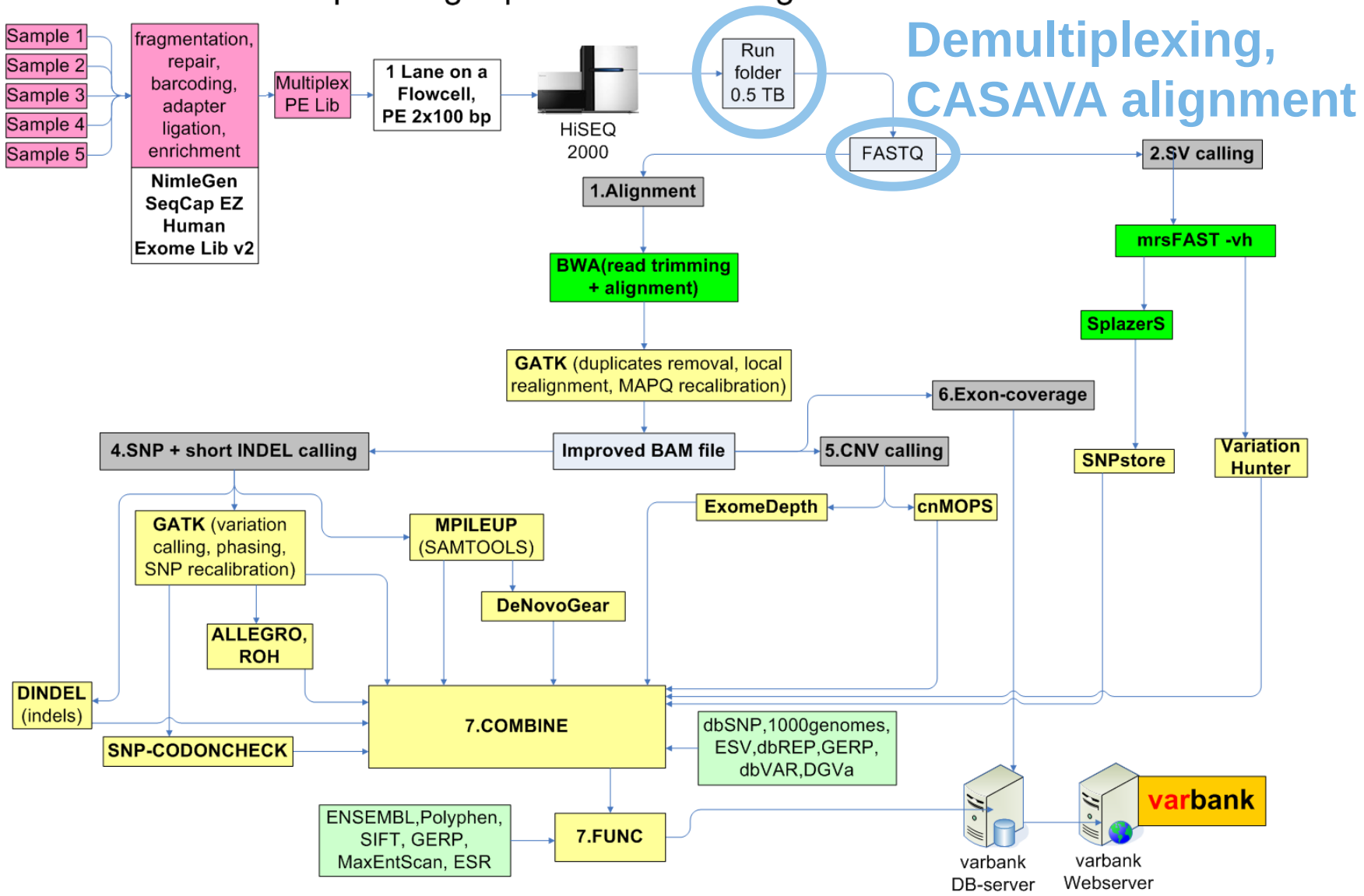


Goals for a new exome analysis pipeline

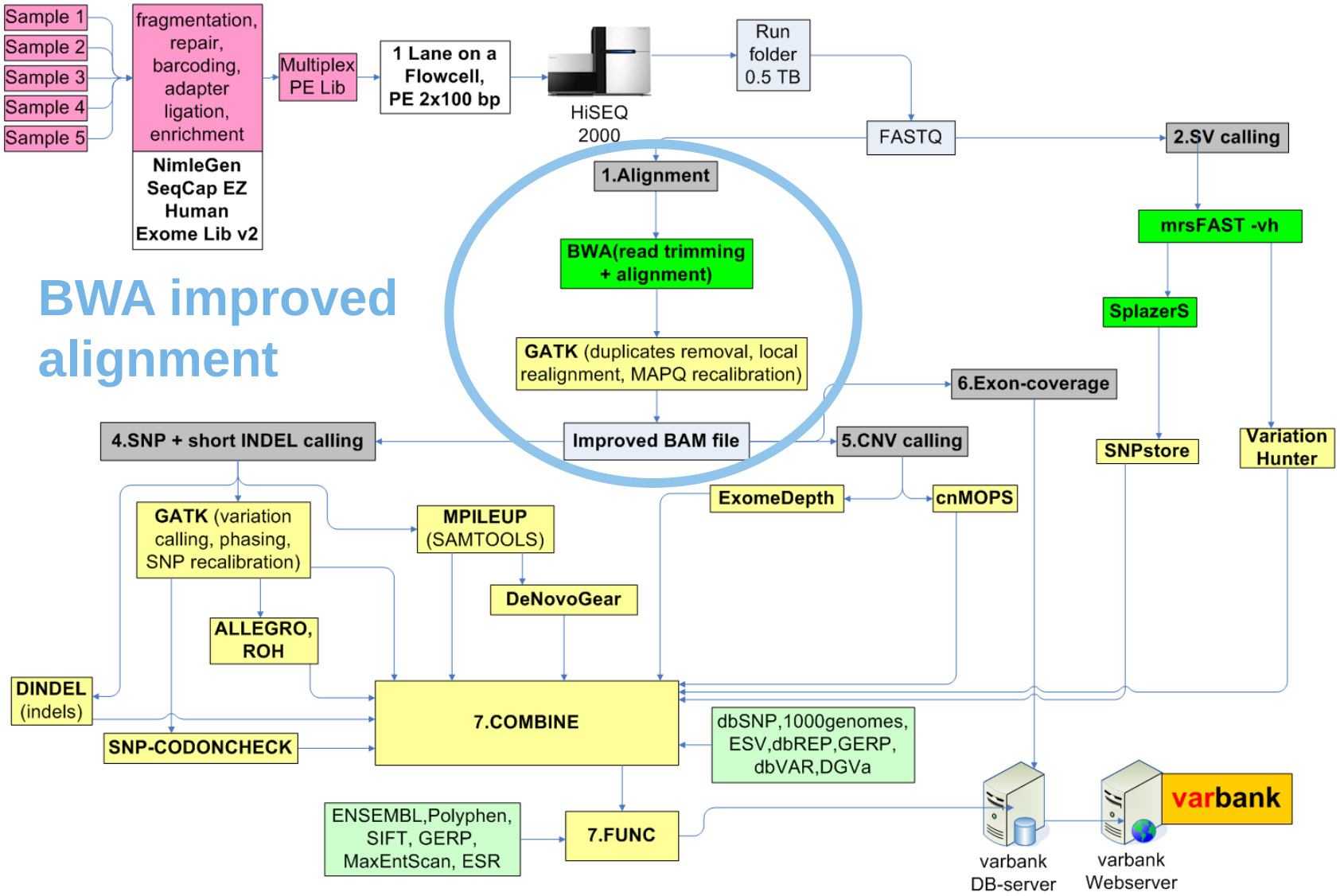
- **Establishing analysis pipeline on HPC Clusters available at the Regional Computing Center Cologne (CHEOPS/SUGI)**
- **Improving alignments**
- **Expanding the range of SNP/short indel callers**
- **Adding support for structural variations and CNVs**
- **Building up a new database and web tools to have remote access to variants, files, statistics, etc.**
- **Design the web tool to make variant browsing, filtering and interpretation as easy as possible**

11/2012
V.2.0.1

Exome Sequencing Pipeline at the Cologne Center for Genomics

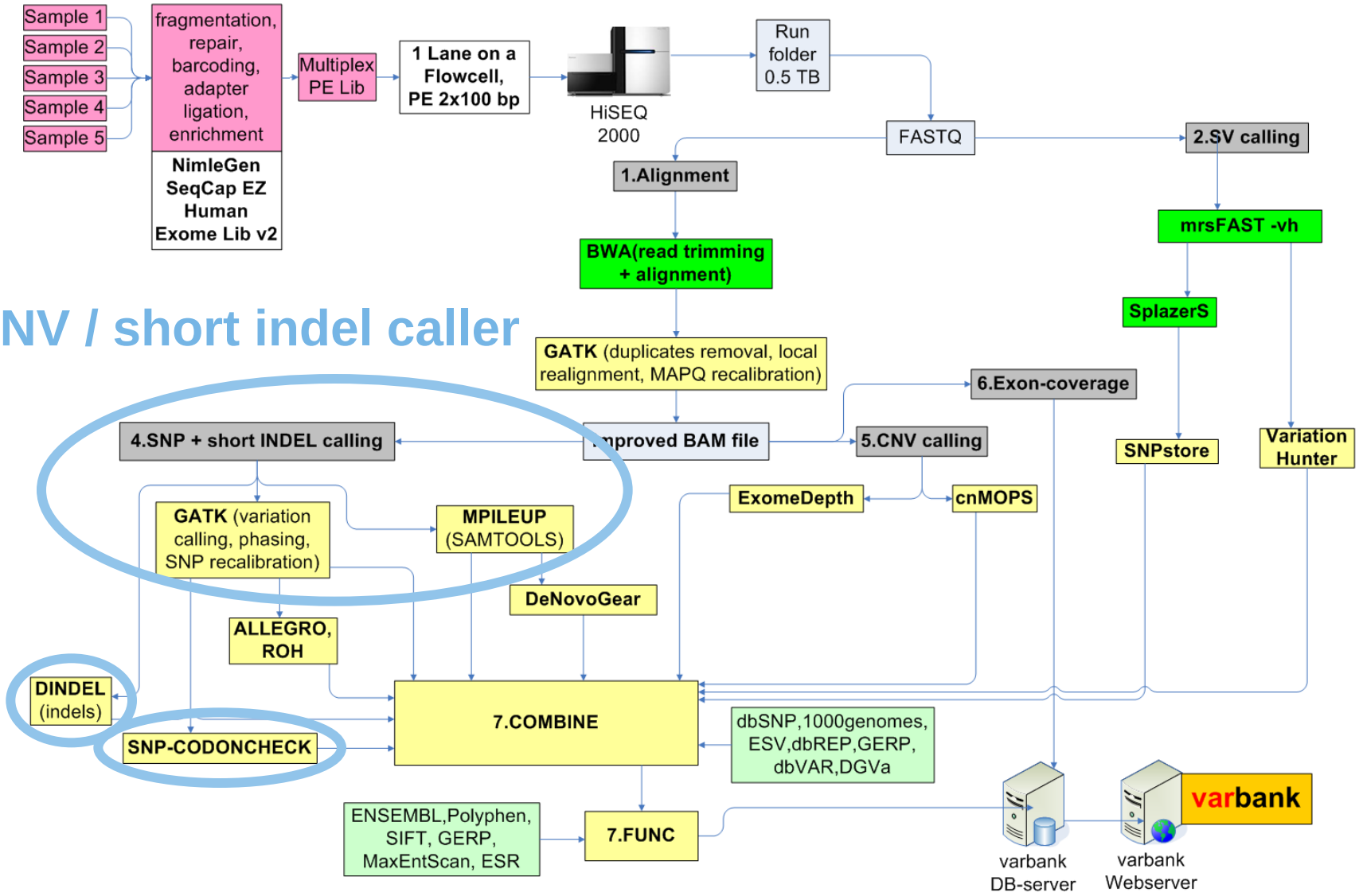


Exome Sequencing Pipeline at the Cologne Center for Genomics



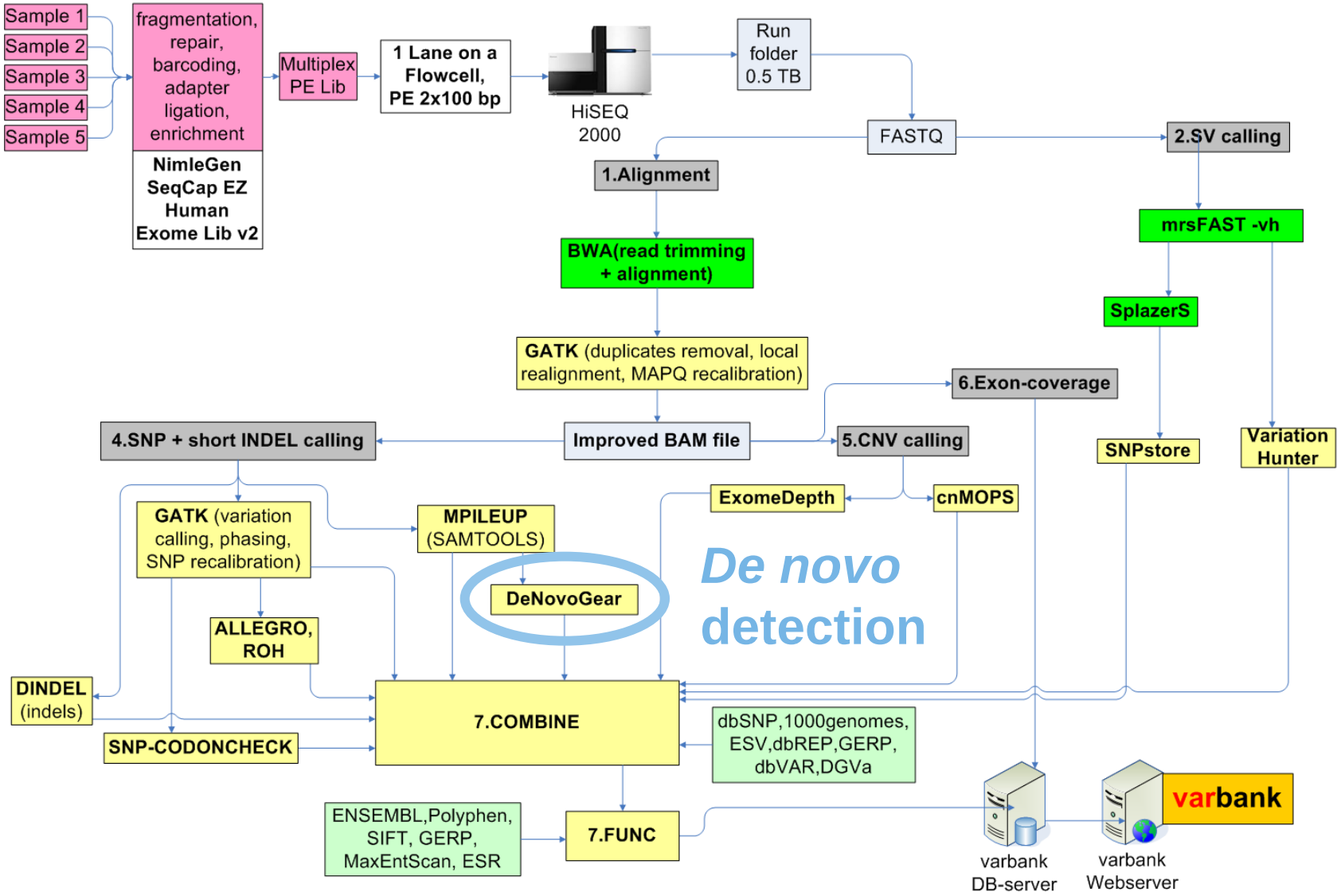
BWA improved alignment

Exome Sequencing Pipeline at the Cologne Center for Genomics

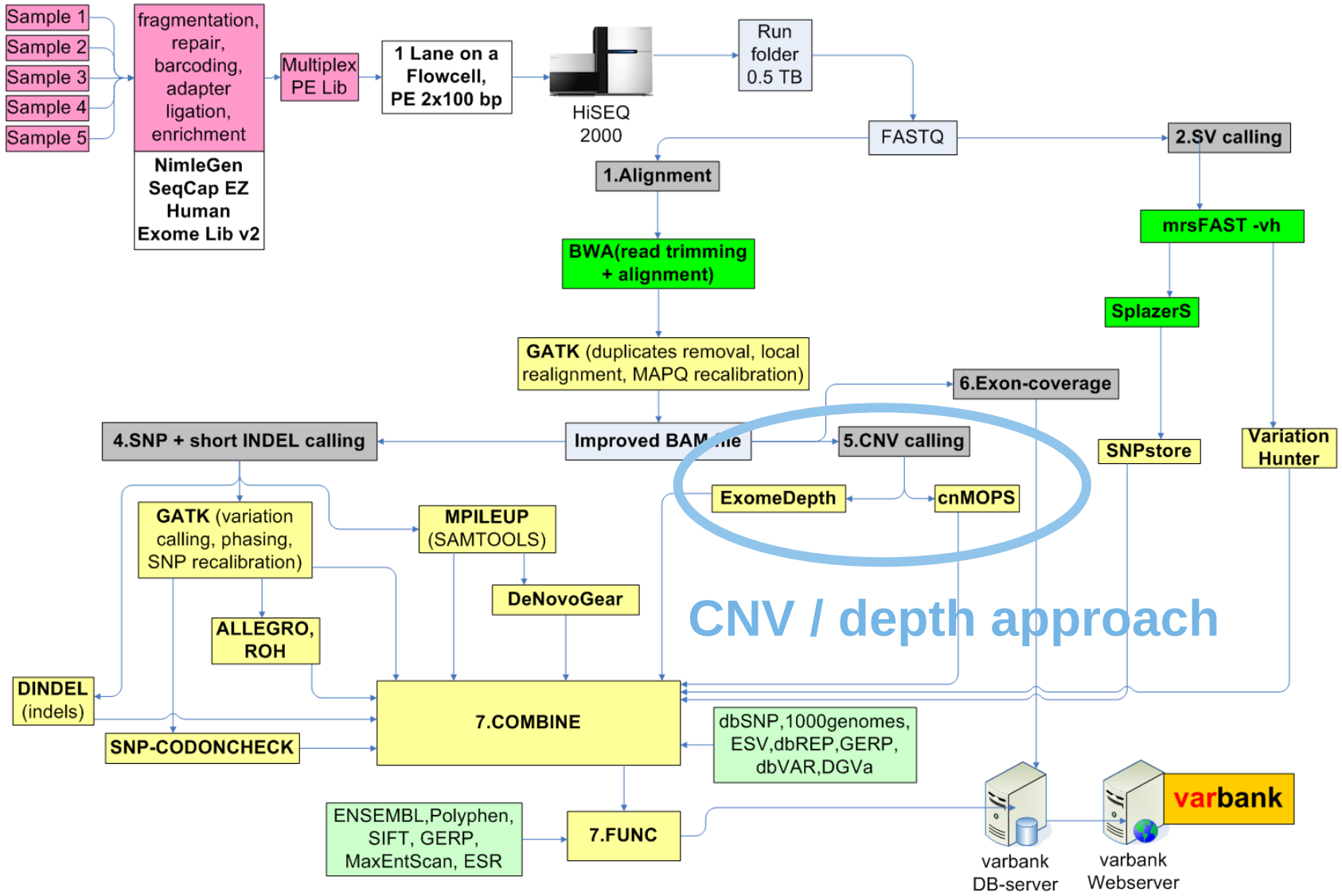


SNV / short indel caller

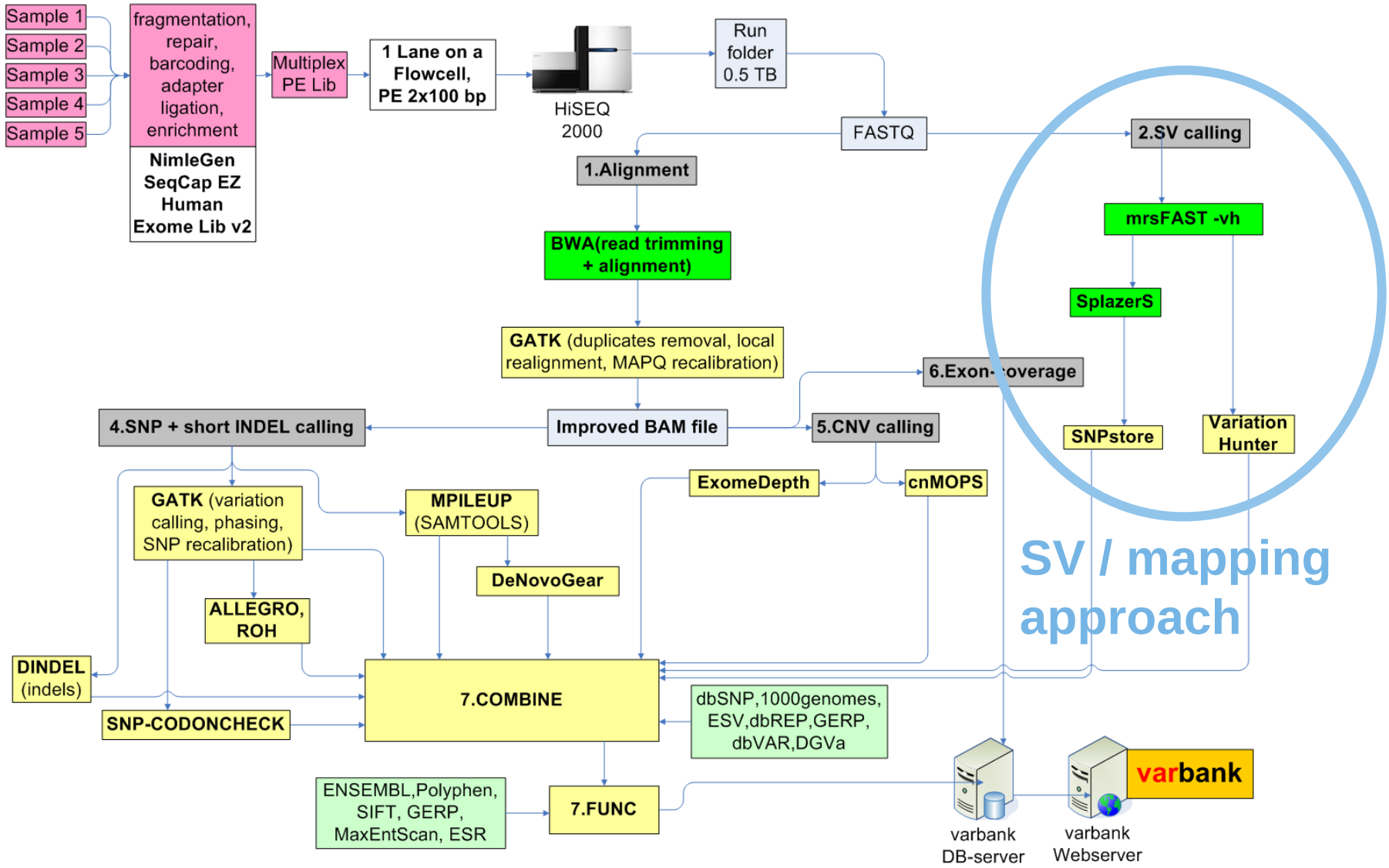
Exome Sequencing Pipeline at the Cologne Center for Genomics



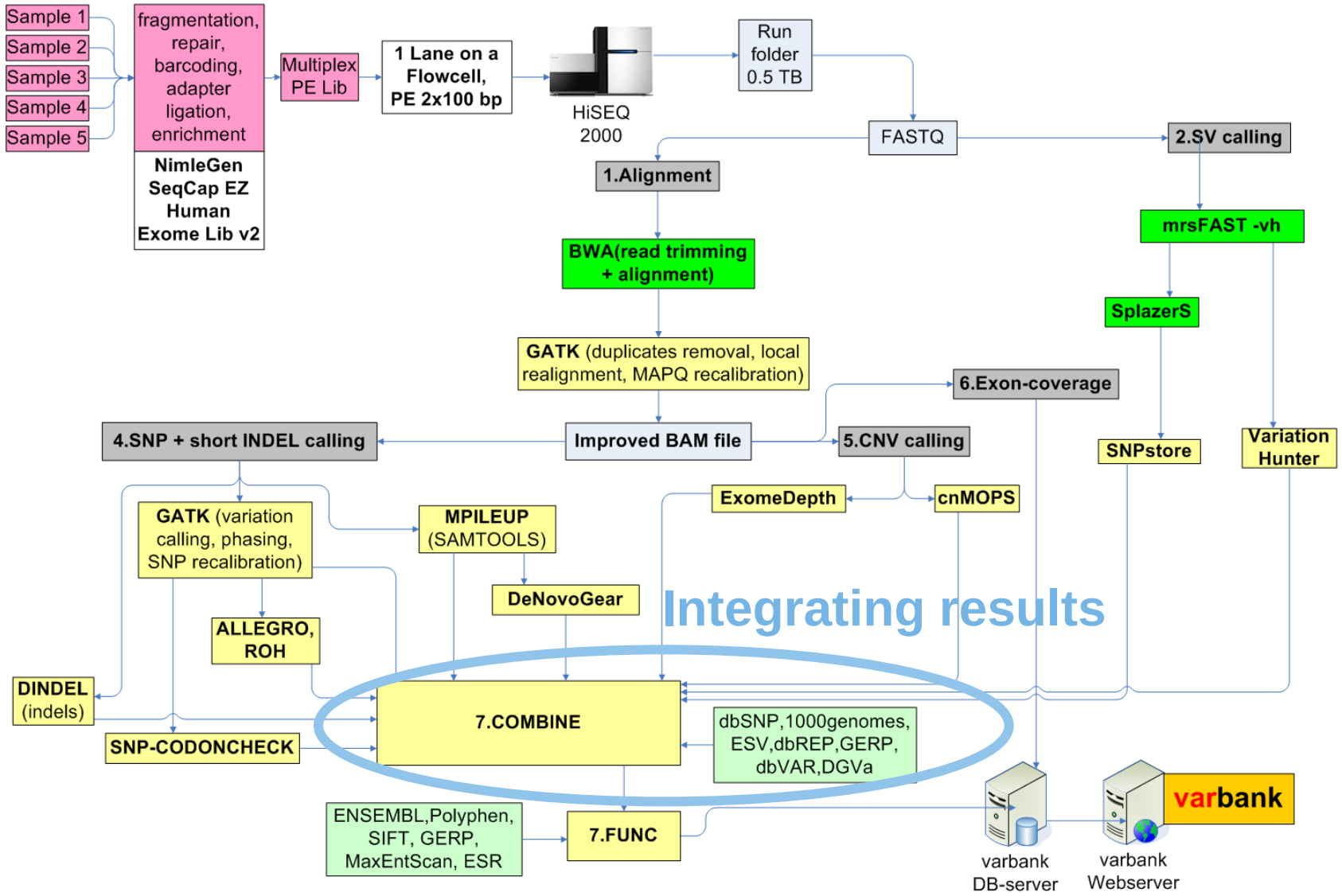
Exome Sequencing Pipeline at the Cologne Center for Genomics



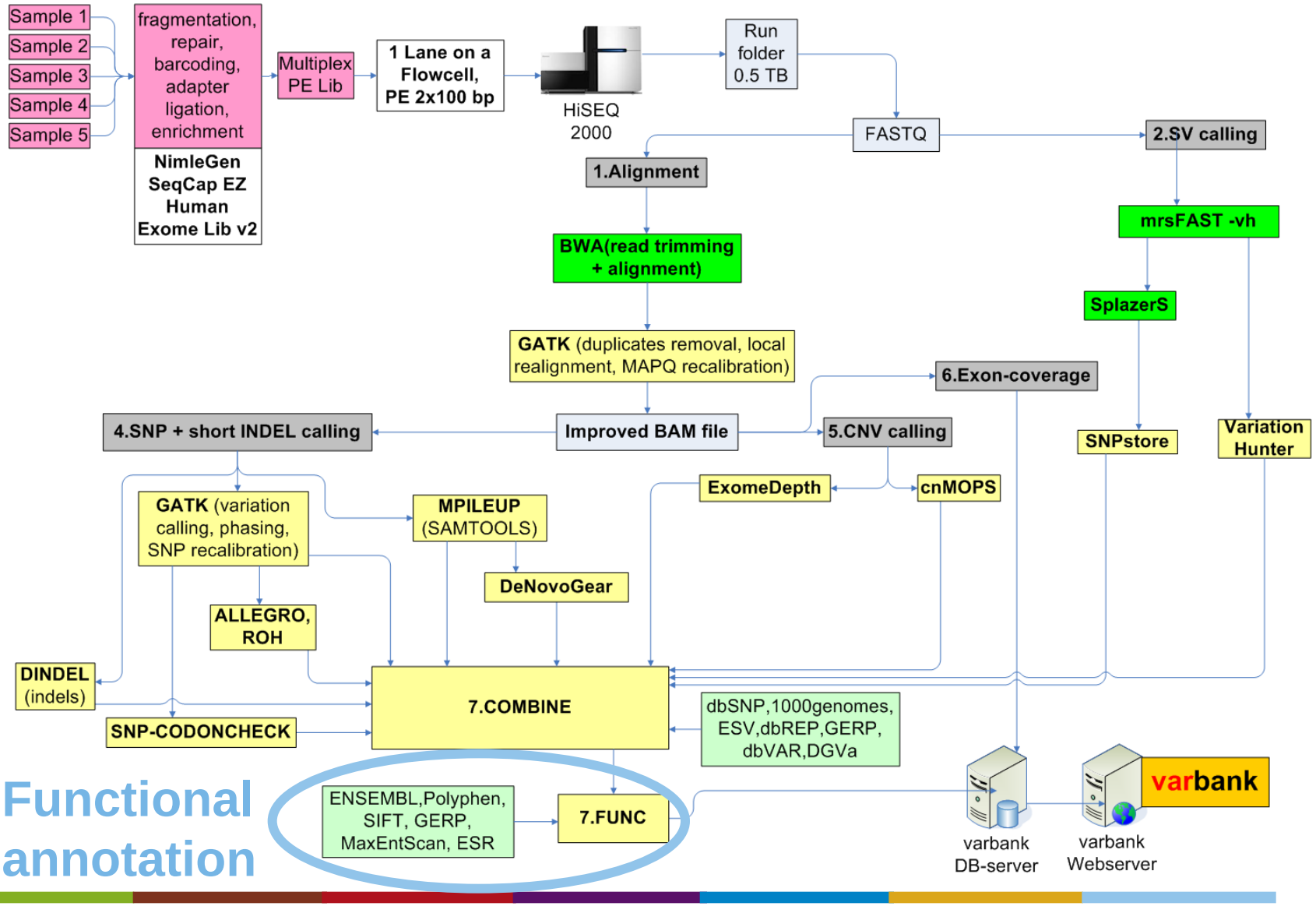
Exome Sequencing Pipeline at the Cologne Center for Genomics



Exome Sequencing Pipeline at the Cologne Center for Genomics



Exome Sequencing Pipeline at the Cologne Center for Genomics



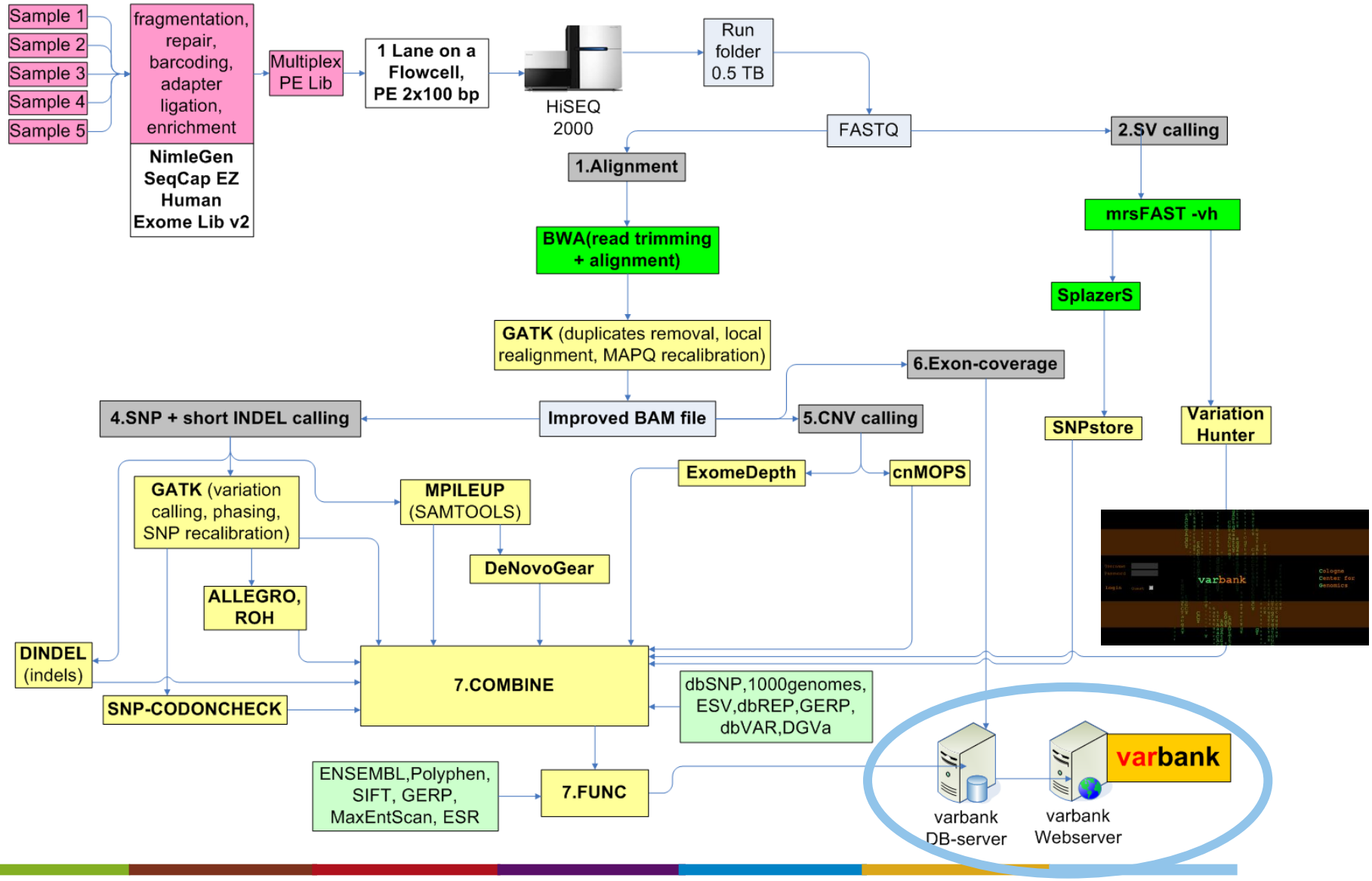
Functional annotation

Cologne Center for Genomics

Universität zu Köln

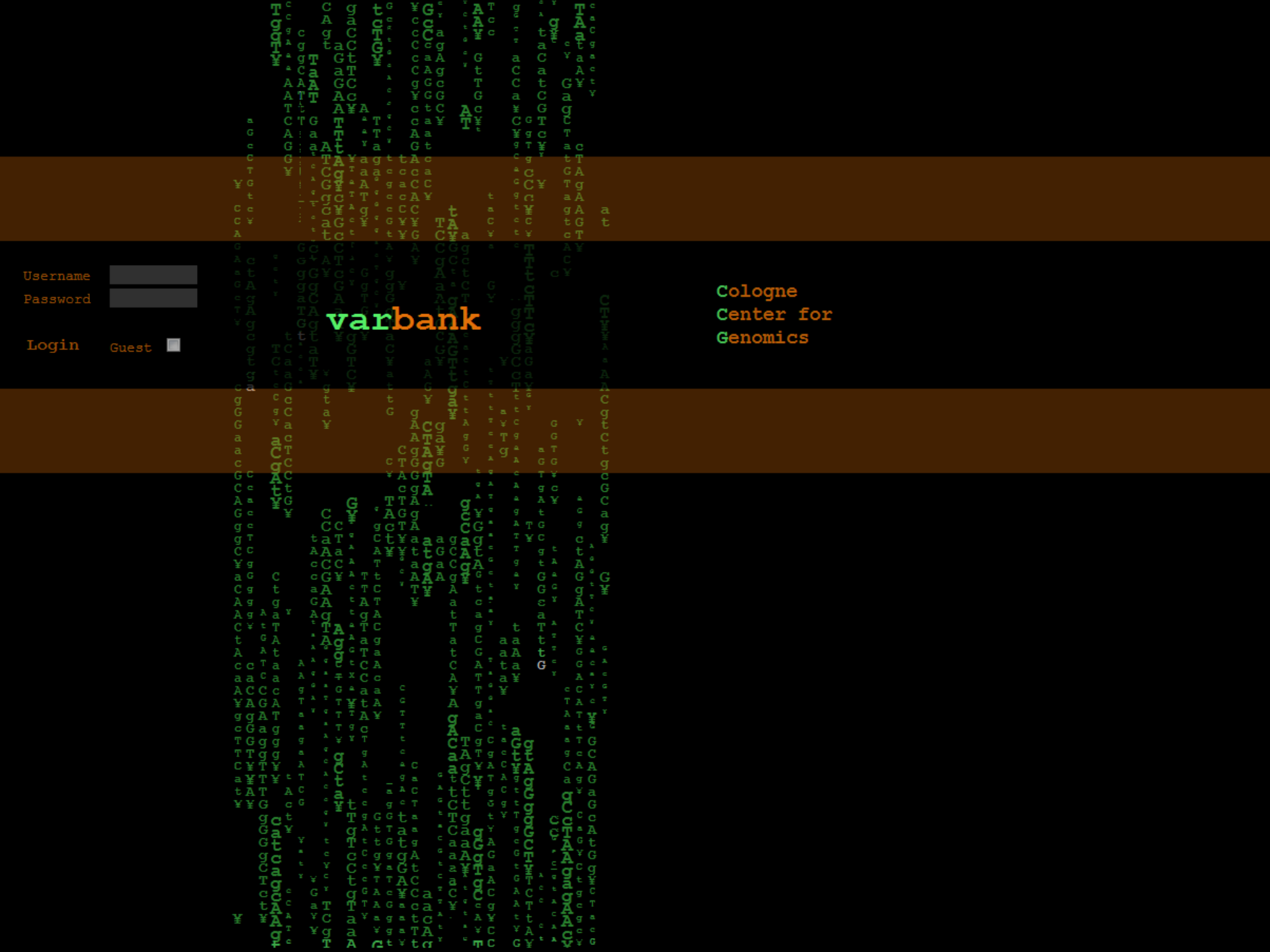
11/2012
V.2.0.1

Exome Sequencing Pipeline at the Cologne Center for Genomics

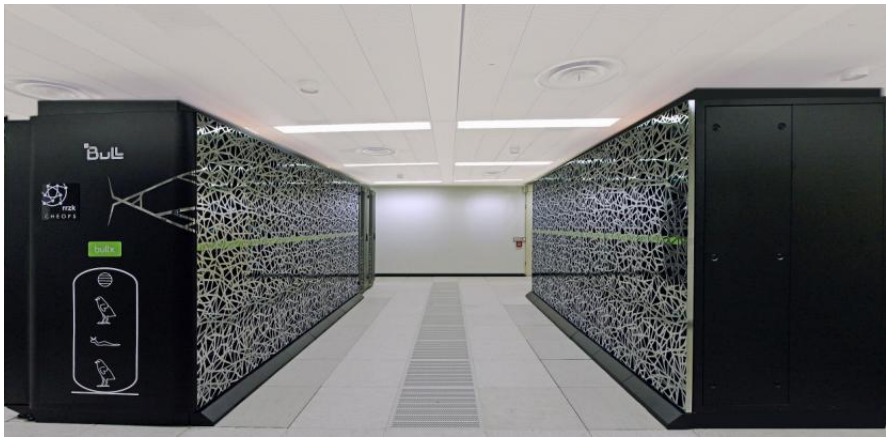


Cologne Center for Genomics

Universität zu Köln


Username
Password
Login
Guest
varbank
Cologne
Center for
Genomics

CHEOPS: HPC@cologne



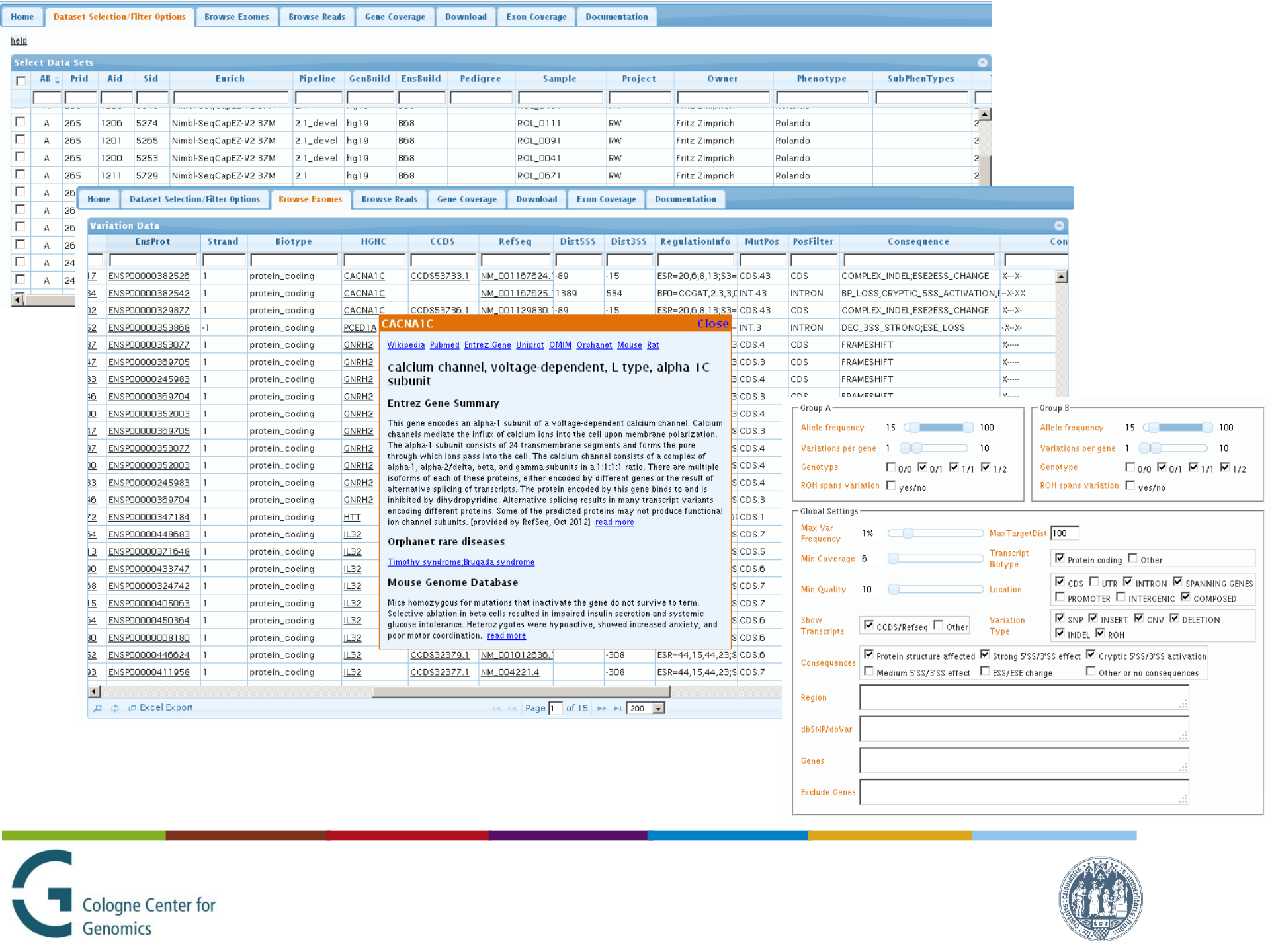
- 100 Tflop/s : in production since 2010
- 210 x 2 Nehalem EP quad-core processors 24 GB RAM
- 432 x 2 Westmere hexa-core processors 24 GB RAM
- 170 x 2 Westmere hexa-core processors 48 GB RAM
- 5 x 2 Nehalem EP quad-core processors 96 GB RAM
- 24 x 4 Nehalem EX octo-core processors 512 GB RAM
- Lustre file system, 500 TB

varbank - features

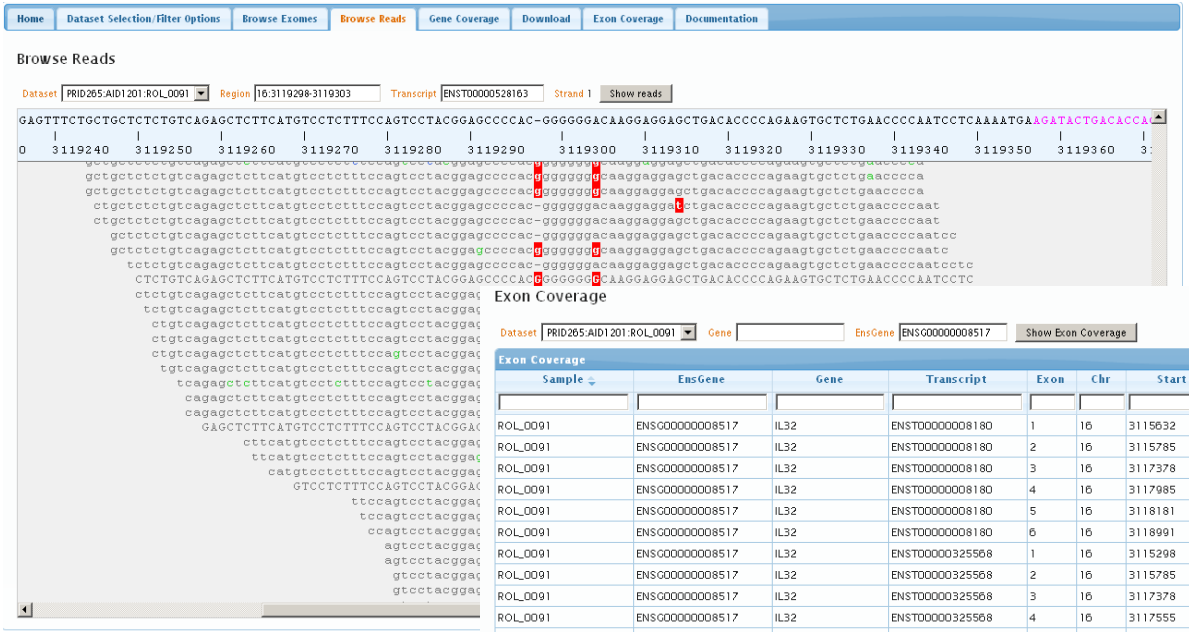
- **Repository for all relevant data files**
- **Fine-tunable access control (login, sample owner, sample sharing)**
- **User-editable for information related to pedigrees and phenotypes**
- **Variant filtering, browsing and exporting**
- **Basic read alignment viewer**
- **Gene coverage viewer**

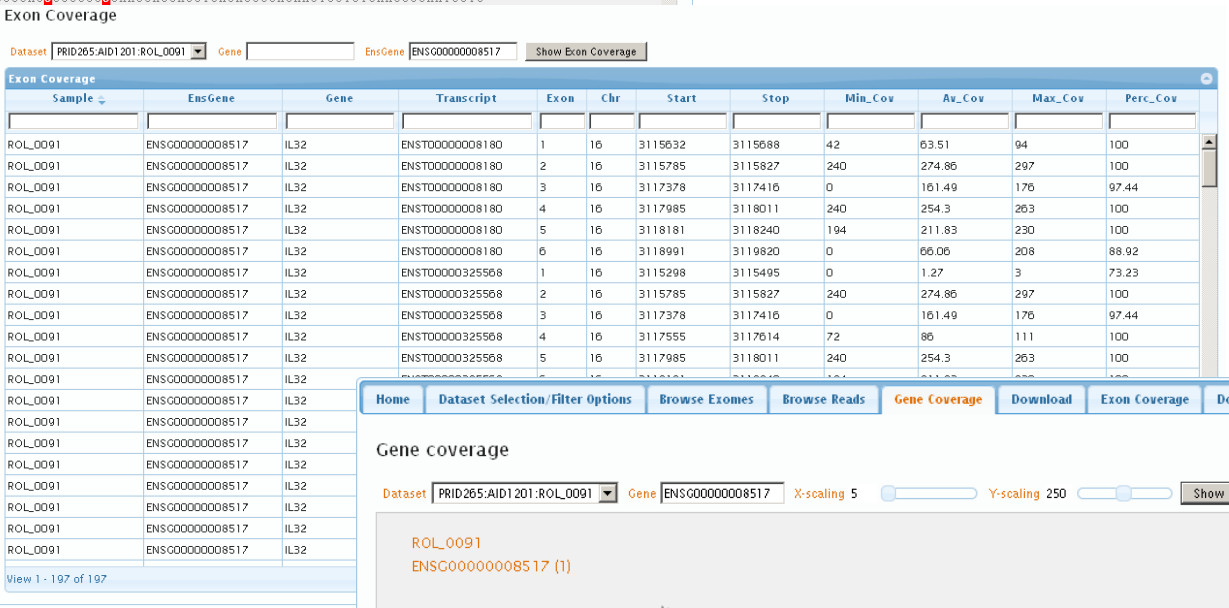
varbank – technical implementation

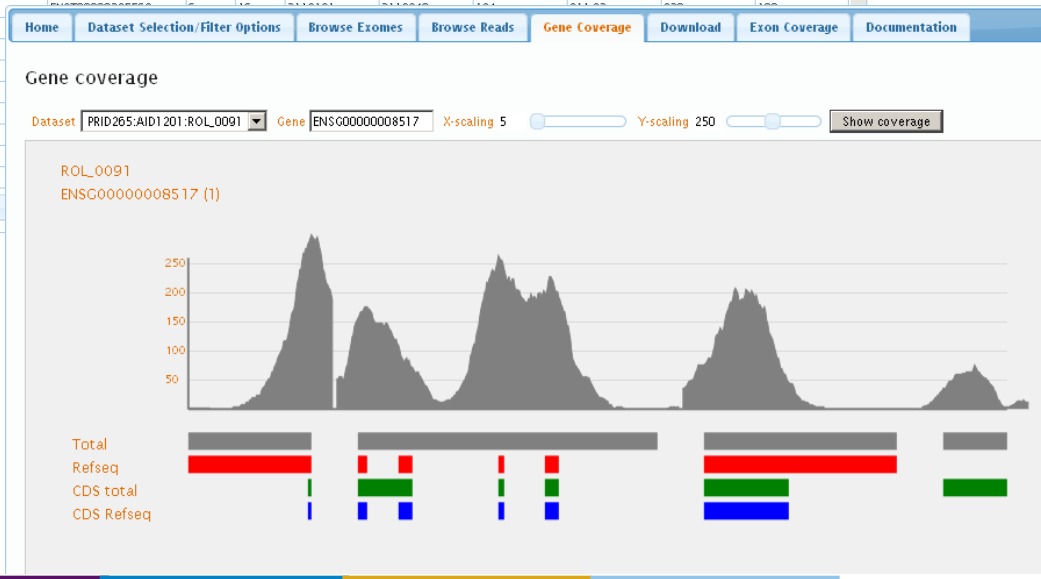
- **Implementation on two VMWare servers located at the Regional Computing Center Cologne (outsourced maintanance, scalable!)**
- **Data base: Oracle 11.2 with table space partitioning**
- **Web server: Apache**
- **Web interface: perl/cgi; java script; jQuery; jQgrid; jQuery.svg; overlibmws; Matrix.js; ~ 5200 code lines**



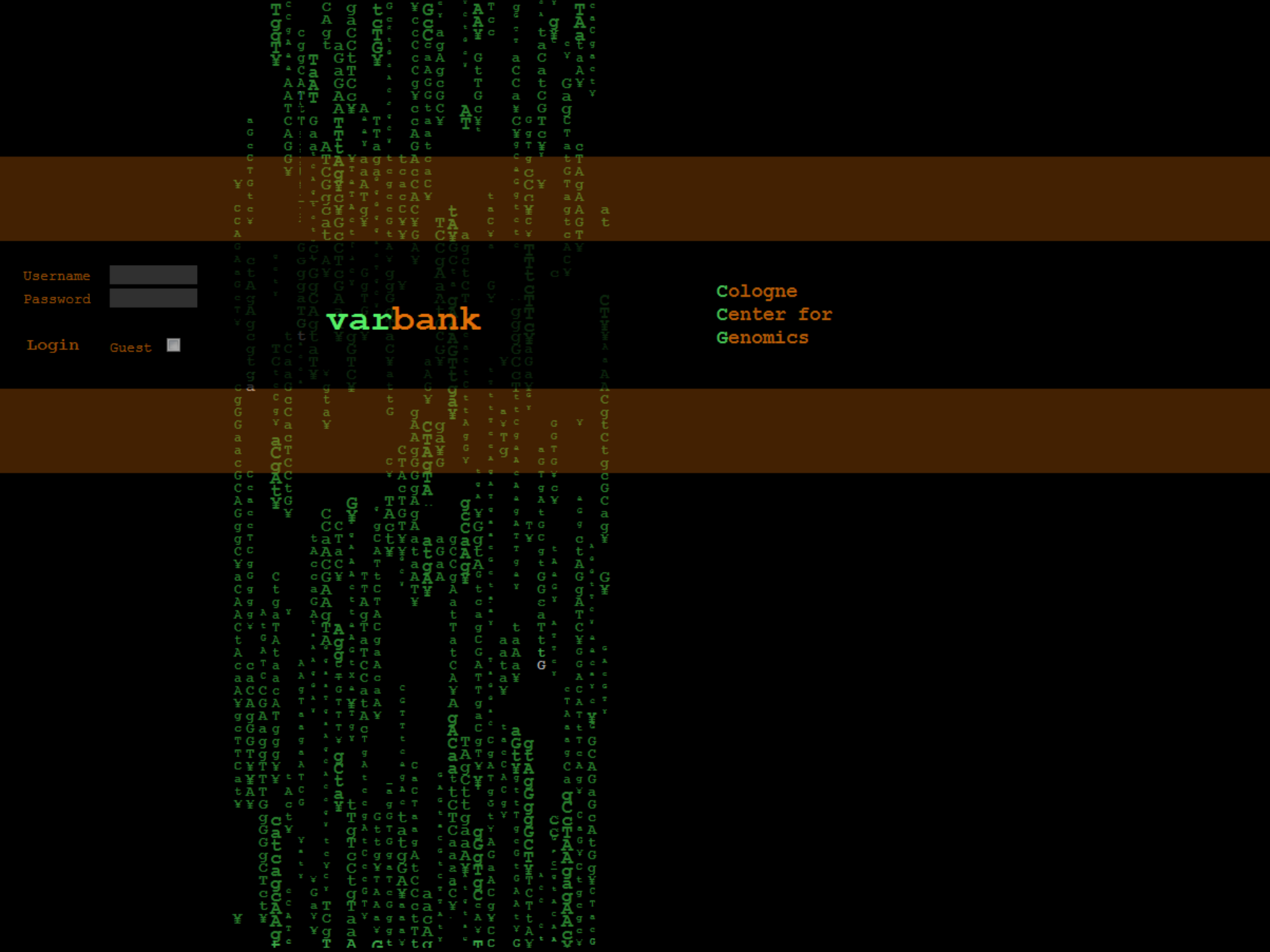
Cologne Center for Genomics







Cologne Center for Genomics


Username
Password
Login
Guest
varbank
Cologne
Center for
Genomics

Conclusion

WES is extremely successful for disease variant identification in Mendelian diseases → clinical geneticists and pediatricians ask for diagnostic exome-seq!

- **few exomes of affected individuals in consanguineous pedigrees**
- **3 or more individuals (affected and unaffected) in dominant families (combination with linkage)**
- **parents/offspring trios for *de-novo* mutations**

CCG's contribution:
>1.500 exomes so far,
(1/3 monogenic, 1/3 cancer, 1/3 complex diseases)

Acknowledgments/Responsibilities

Lab & Planing

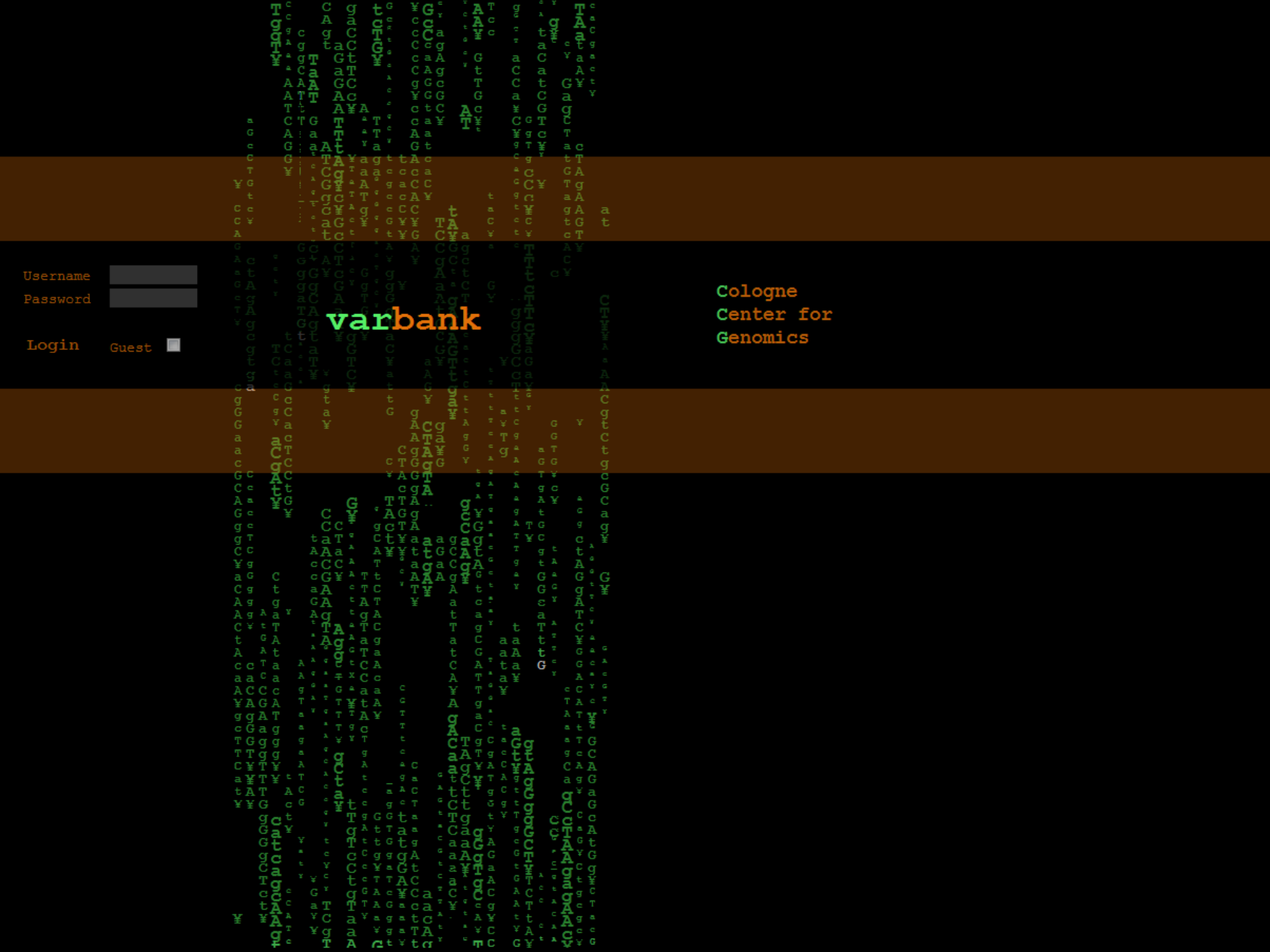
- **Janine Altmüller**
  (project managment)
- **Christian Becker**
  (library prep, sequencing)
- **Elisabeth Kirst**
  (library prep, sequencing)
- **Marek Franitza**
  (library prep, sequencing)

Computing Center

- **Viktor Achter**
  (HPC administration)
- **Ulrich Lang**
  (resource management)

Bioinformatics

- **Susanne Motameny**
  (HPC exome pipeline, CNV)
- **Peter Frommolt**
  (HPC, NGSrich, Dindel, Tumor/Transcriptome analysis)
- **Kamel Jabbari**
  (SV, CoGIE project data analysis)
- **Amit Kawala**
  (BWA Alignment)
- **Holger Thiele**
  (pipeline planning, data integration, functional annotation, trios, database, web front end)
- **Wilfried Gunia**
  (server administration, deNovo@HPC)


Username
Password
Login
Guest
varbank
Cologne
Center for
Genomics