

Exome sequencing at the CCG

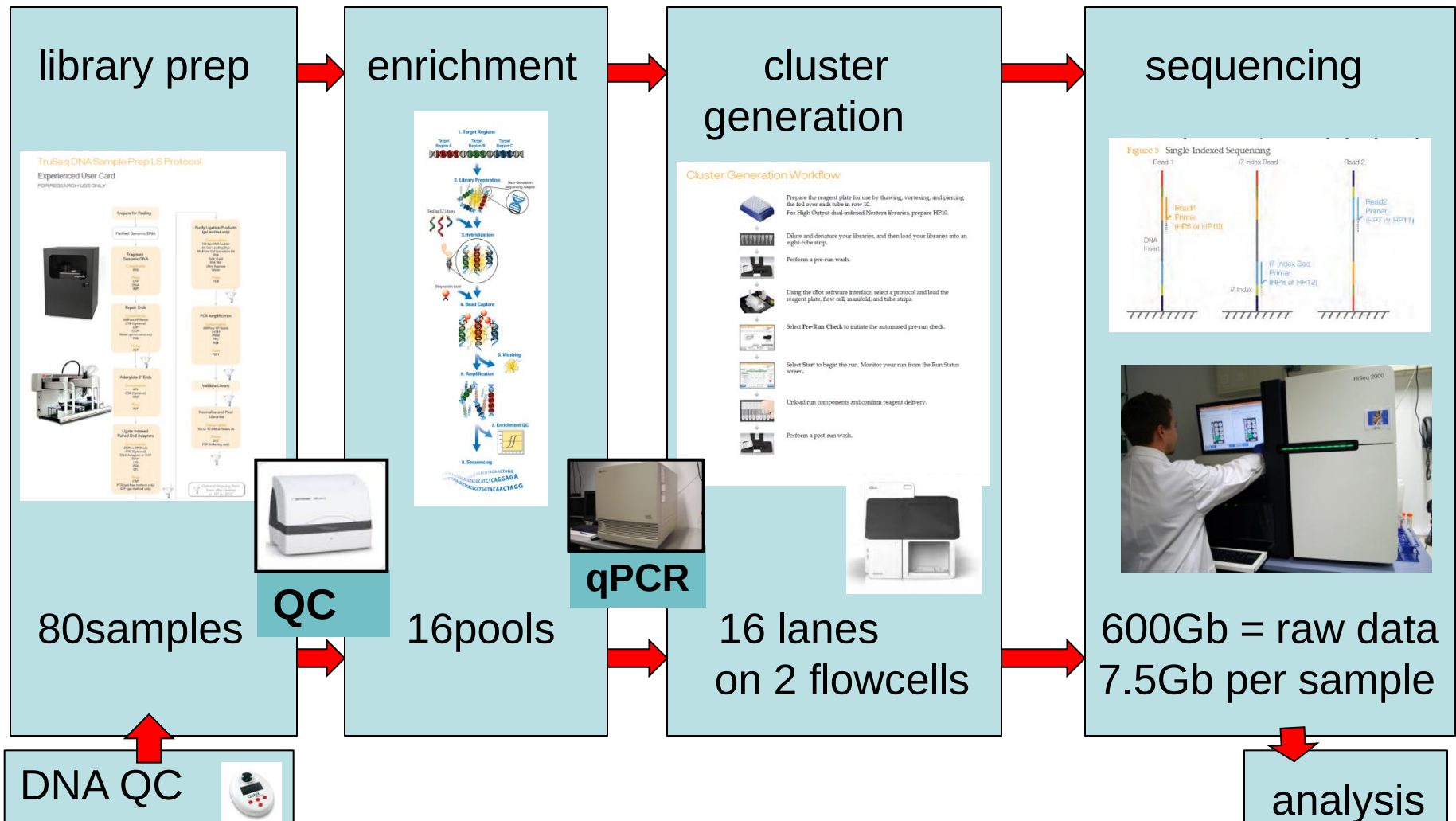
“NGS an der Schnittstelle von Grundlagen- und translationaler Forschung”



Peter Nürnberg

**TMF-Workshop
Berlin, 7.12.2012**

The exome sequencing pipeline



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

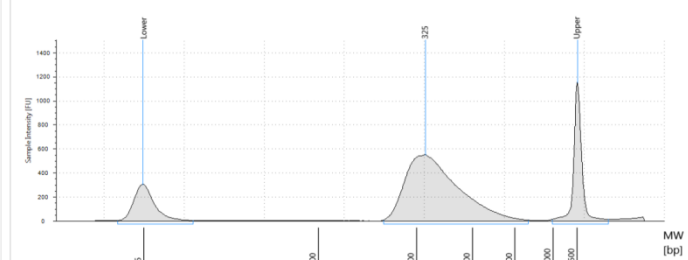
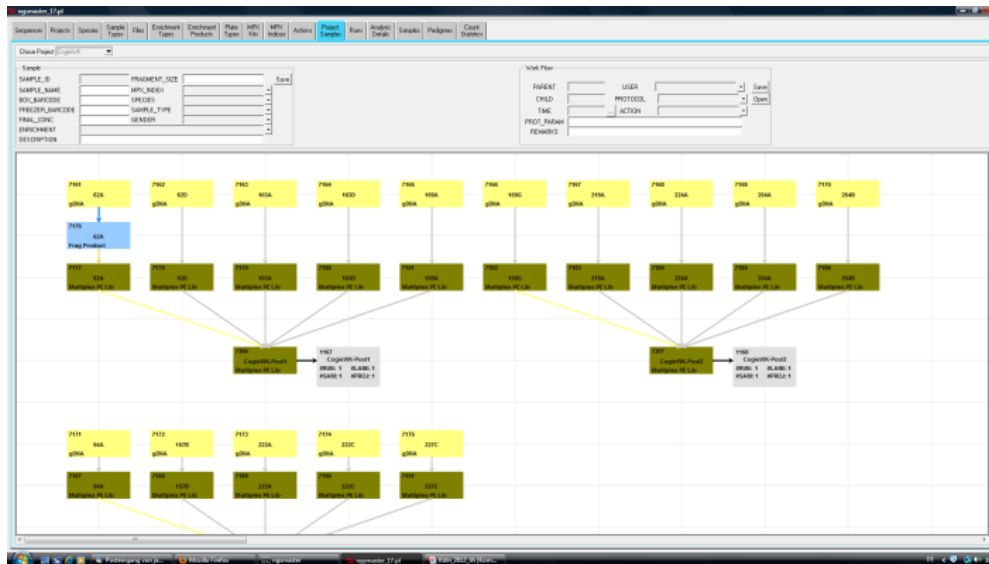
illumina

TruSeq® DNA
Sample Preparation Guide



TruSeq

FOR RESEARCH USE ONLY
ILLUMINA PROPRIETARY
Part # 1502486 Rev. C
July 2012



G01_Lib

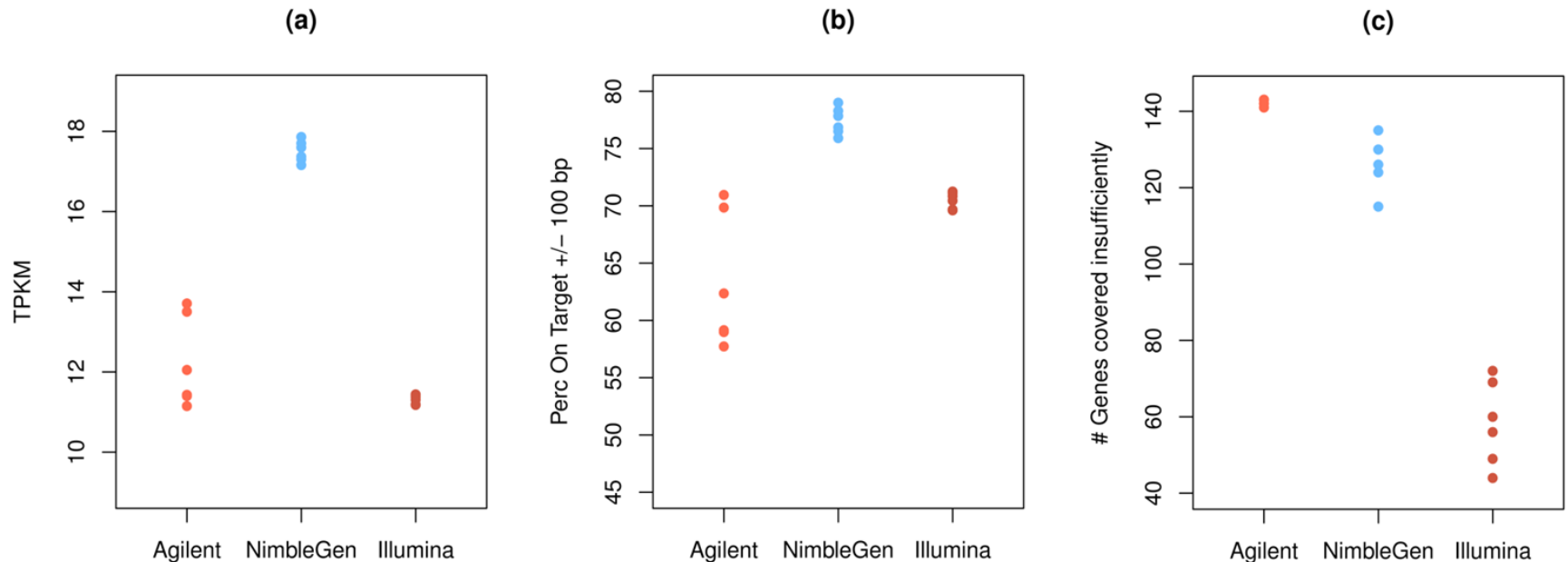
Wavelength	MW [bp]	Conc. [ng/μl]	Molarity [nmol/l]	% of Integrated Area	Peak Comment	Observations
Sam ple	25	4.46	271			Lower Marker
Sam ple	325	20.8	96.9	100.00		
Sam ple	1,500	6.50	6.57			Upper Marker

Enrichment: different products

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

Assessing the enrichment performance in targeted resequencing experiments

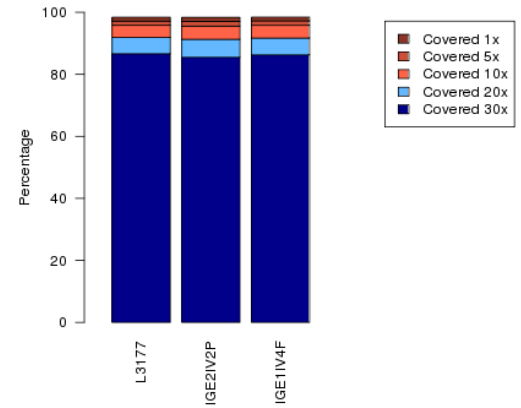
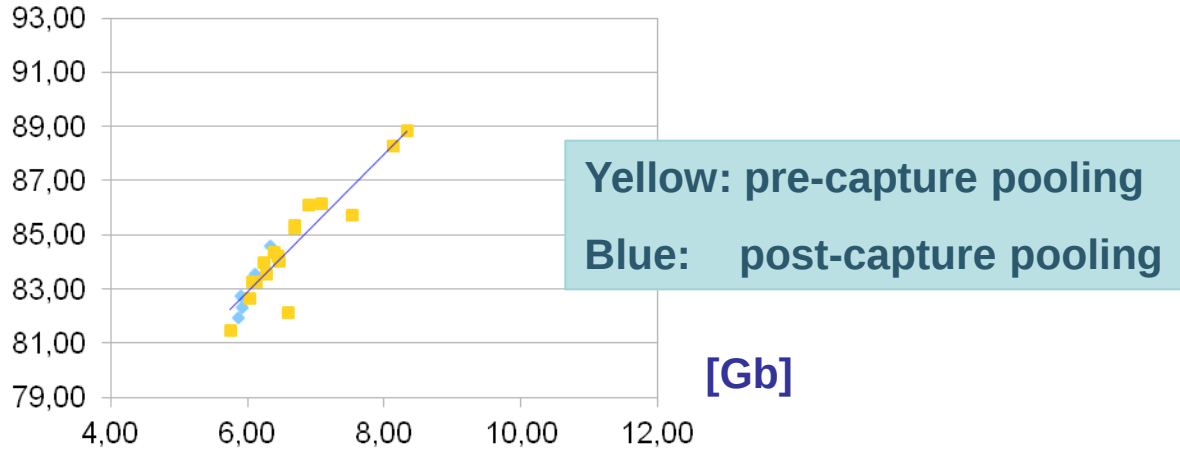
Peter Frommolt et al., Hum Mut 2011



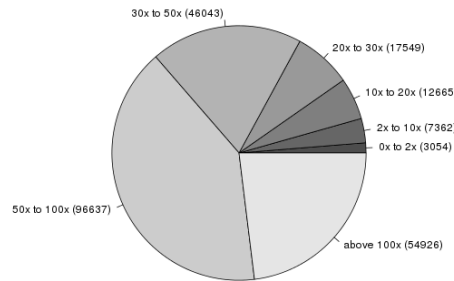
TPKM = On-Target reads Per Kilobase target region and Million mappable reads

Nimblegen v2: strategy considerations

% of target > 30X

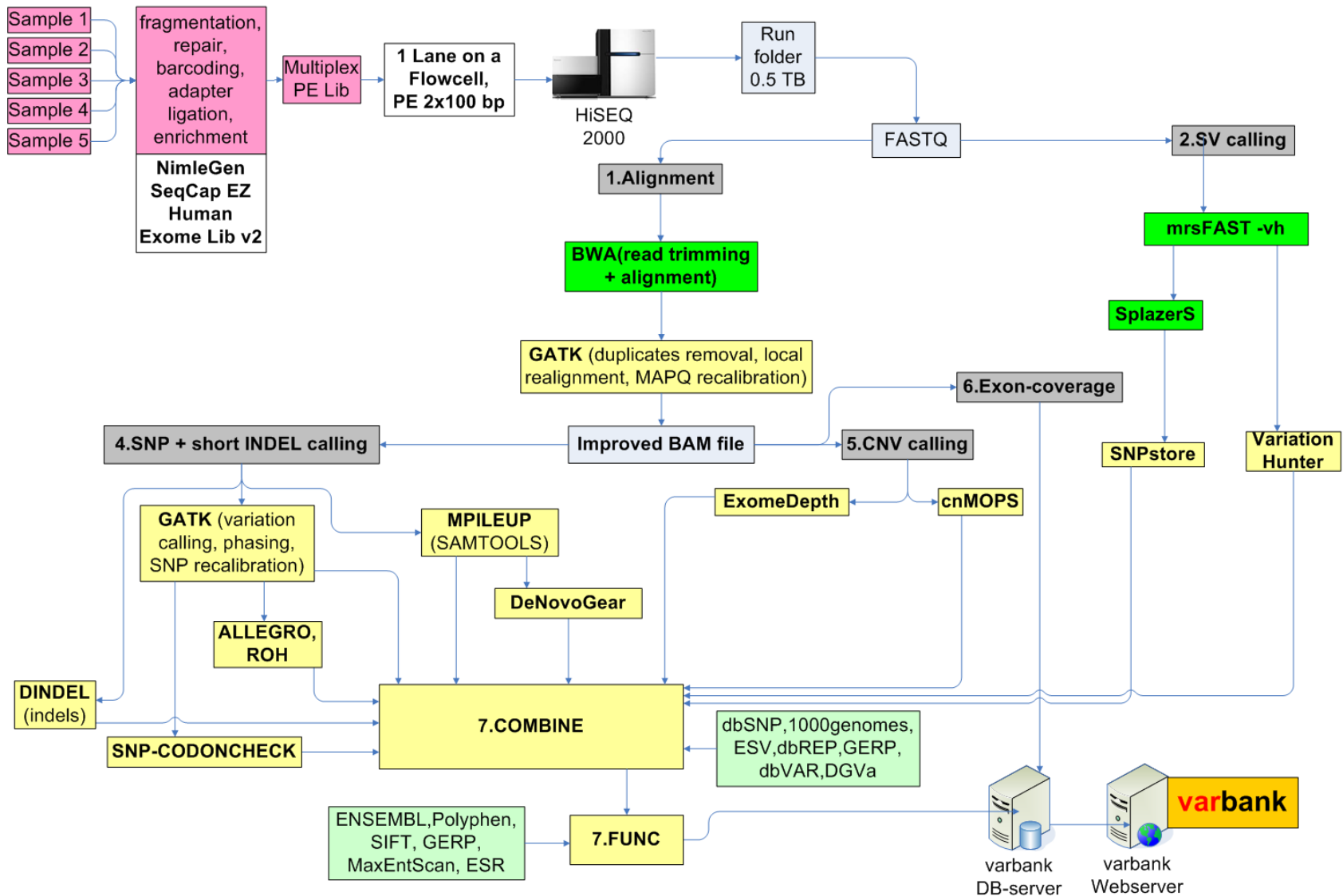


# Reads	64620968
# On Target $\pm$ 100 bp	44549948
Target Size (bp)	44234111
# Target Regions	238236
Coverage Mean	77.27
Coverage Std Dev	64.9
Covered 1x	97.93%
Covered 5x	96.46%
Covered 10x	94.61%
Covered 20x	89.55%
Covered 30x	82.45%
TPKM	15.59



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

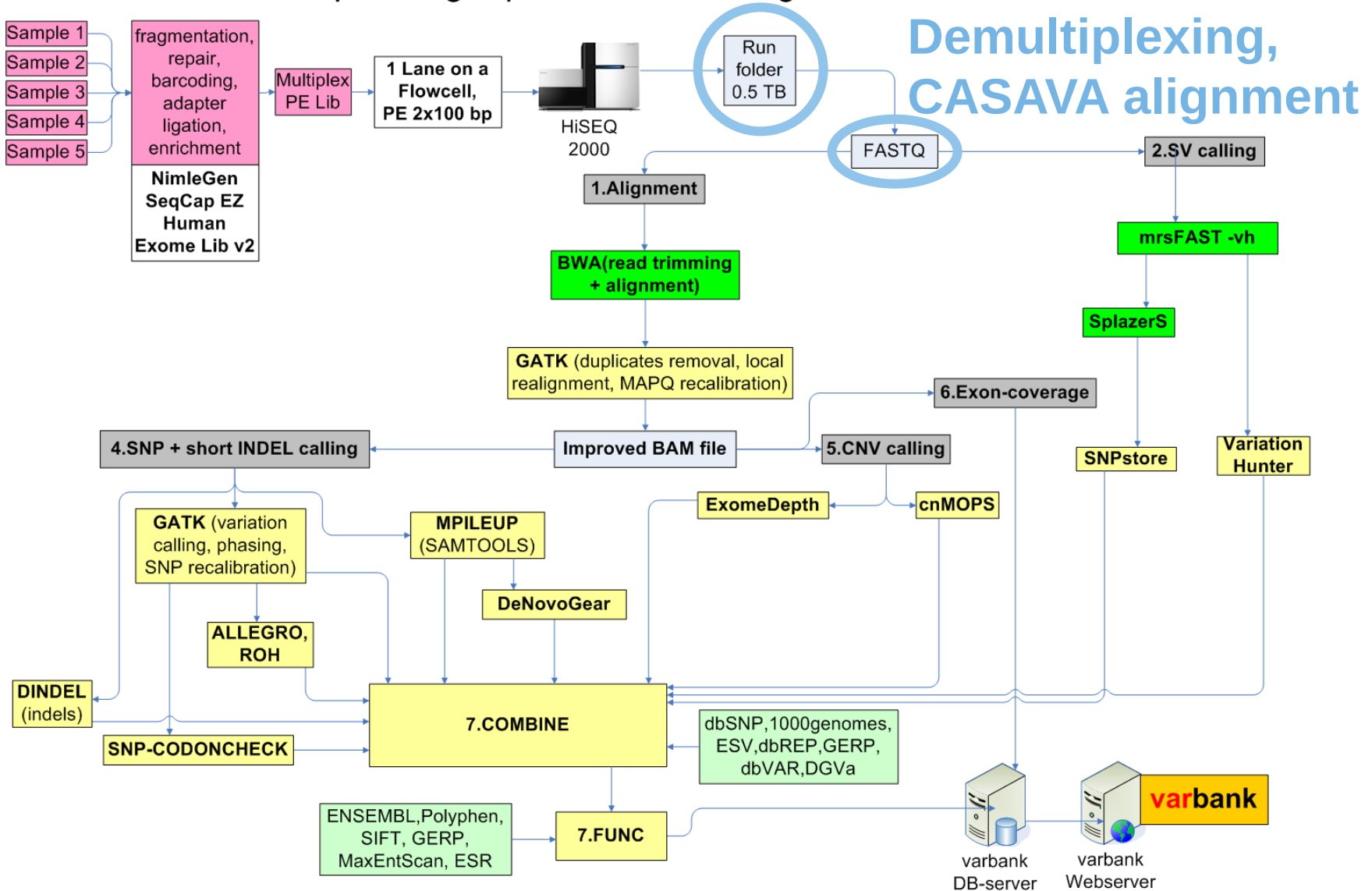
Exome Sequencing Pipeline at the Cologne Center for Genomics



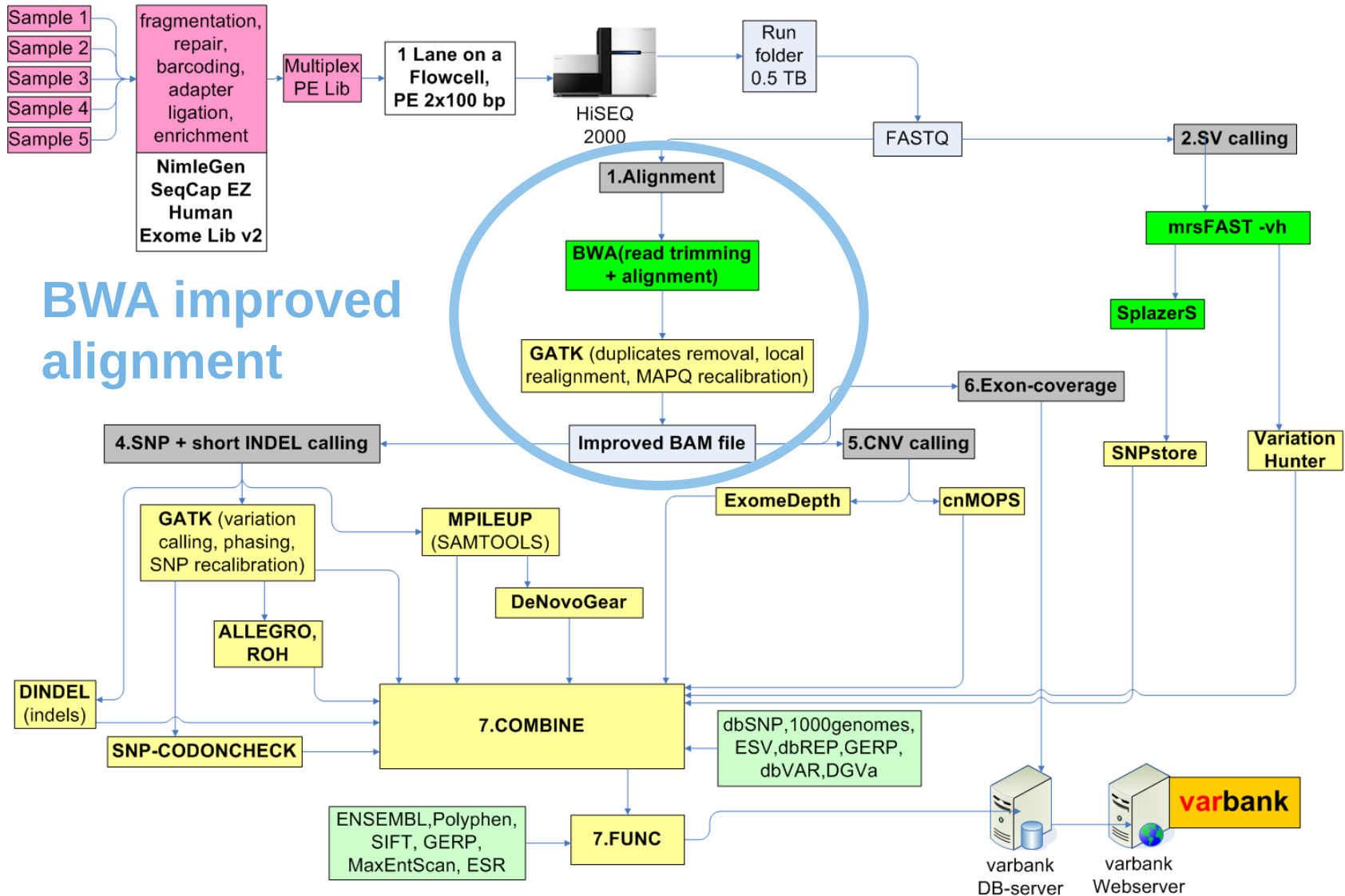
Goals for a new exome analysis pipeline

- 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

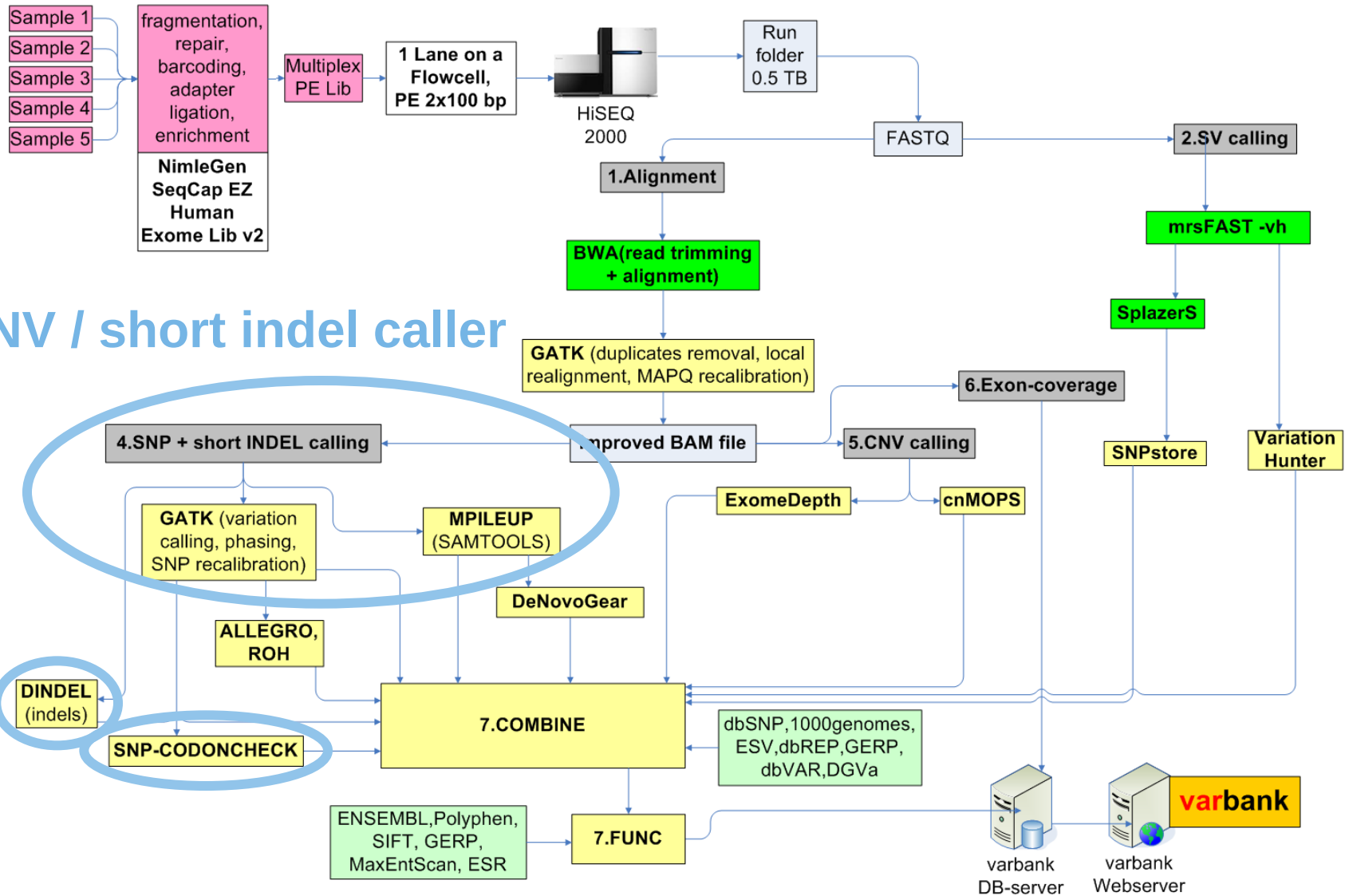
Exome Sequencing Pipeline at the Cologne Center for Genomics



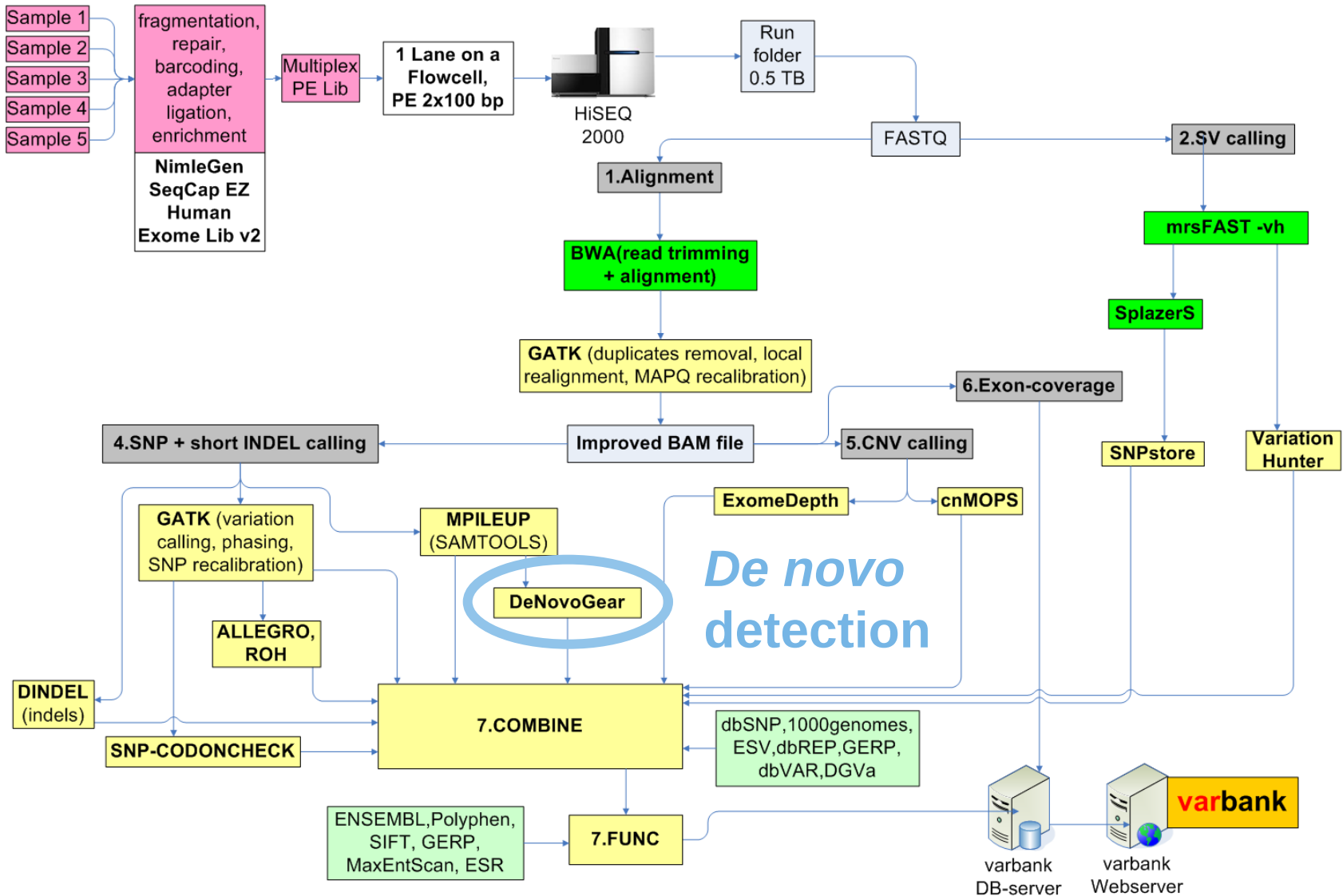
Exome Sequencing Pipeline at the Cologne Center for Genomics



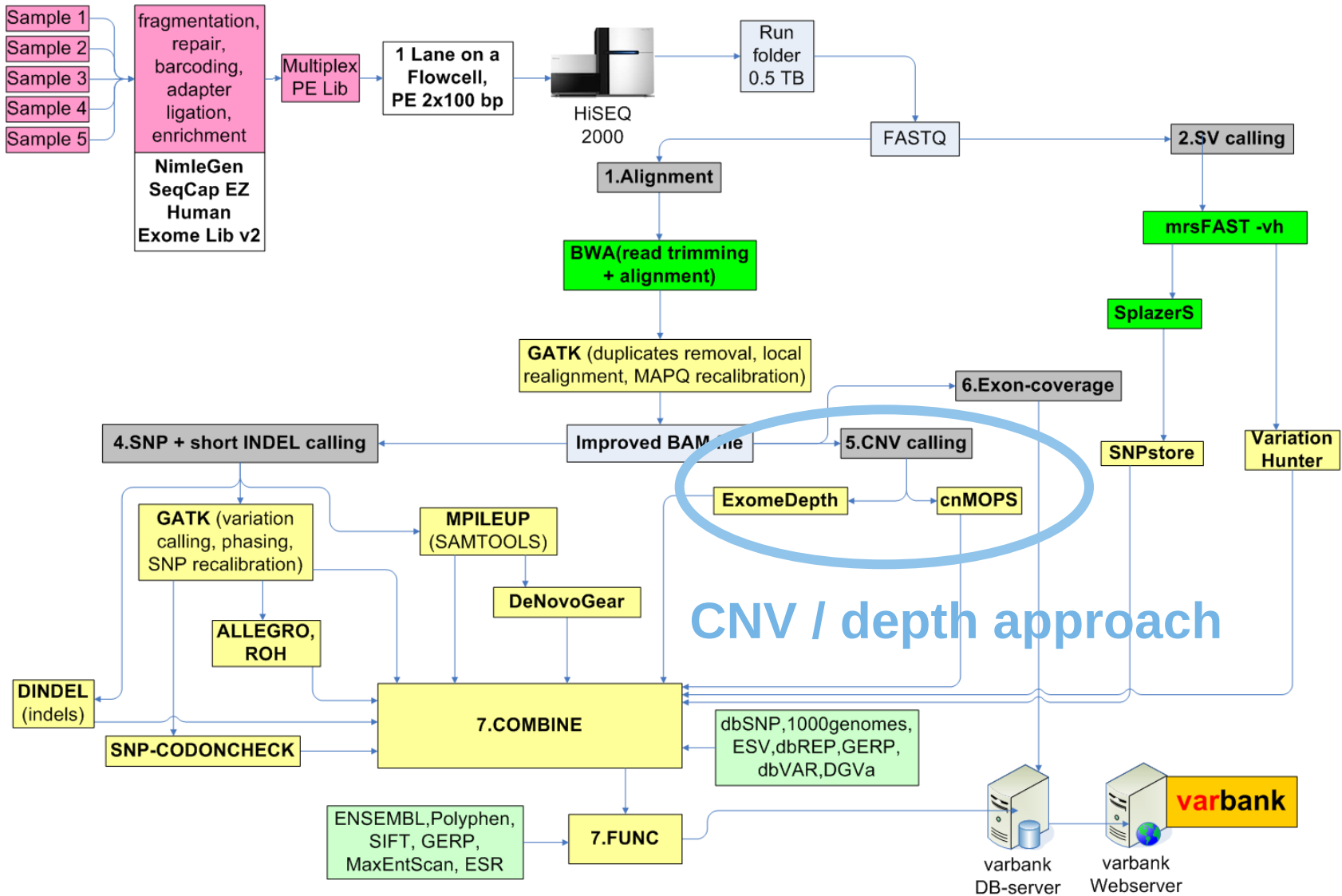
Exome Sequencing Pipeline at the Cologne Center for Genomics



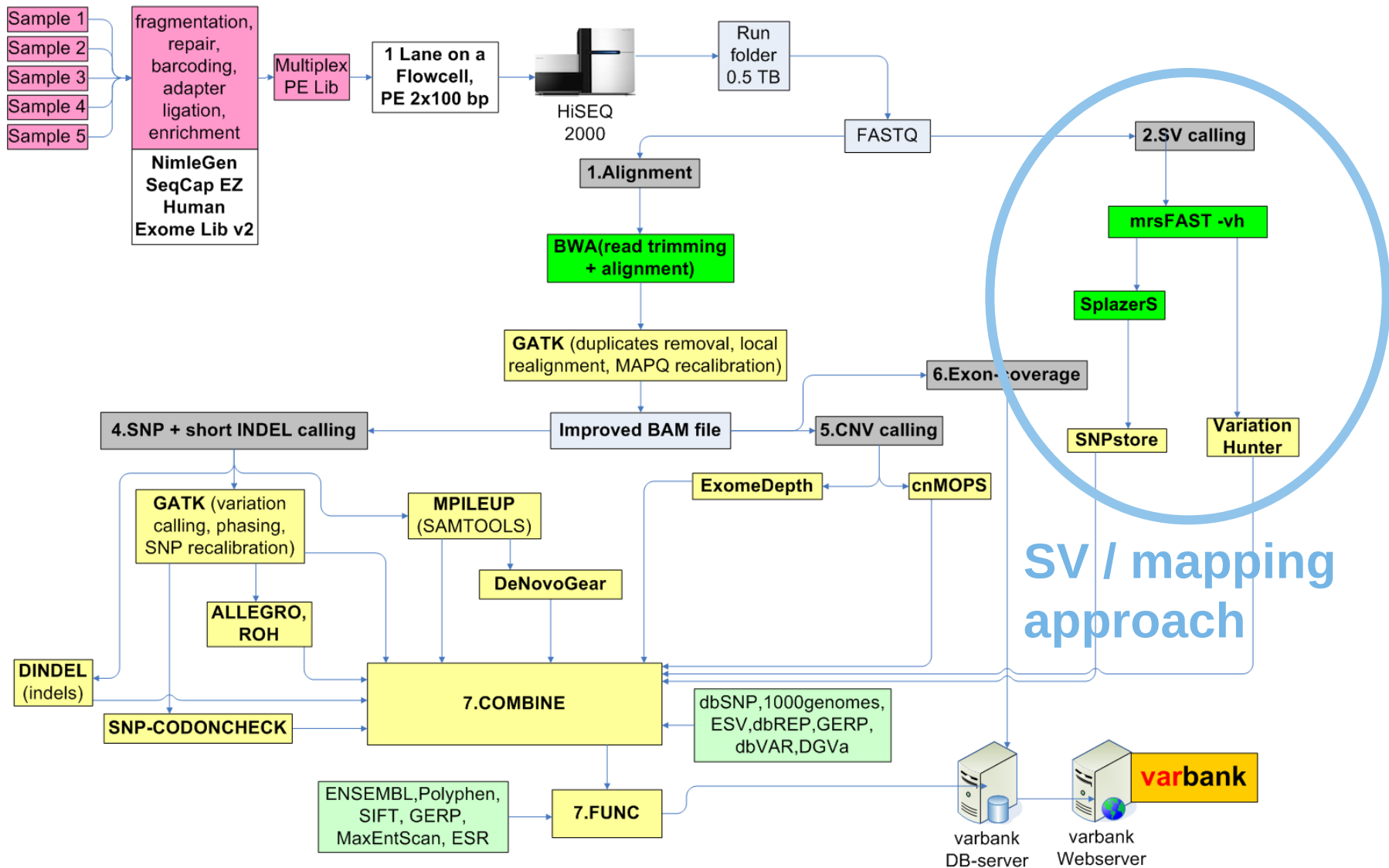
Exome Sequencing Pipeline at the Cologne Center for Genomics



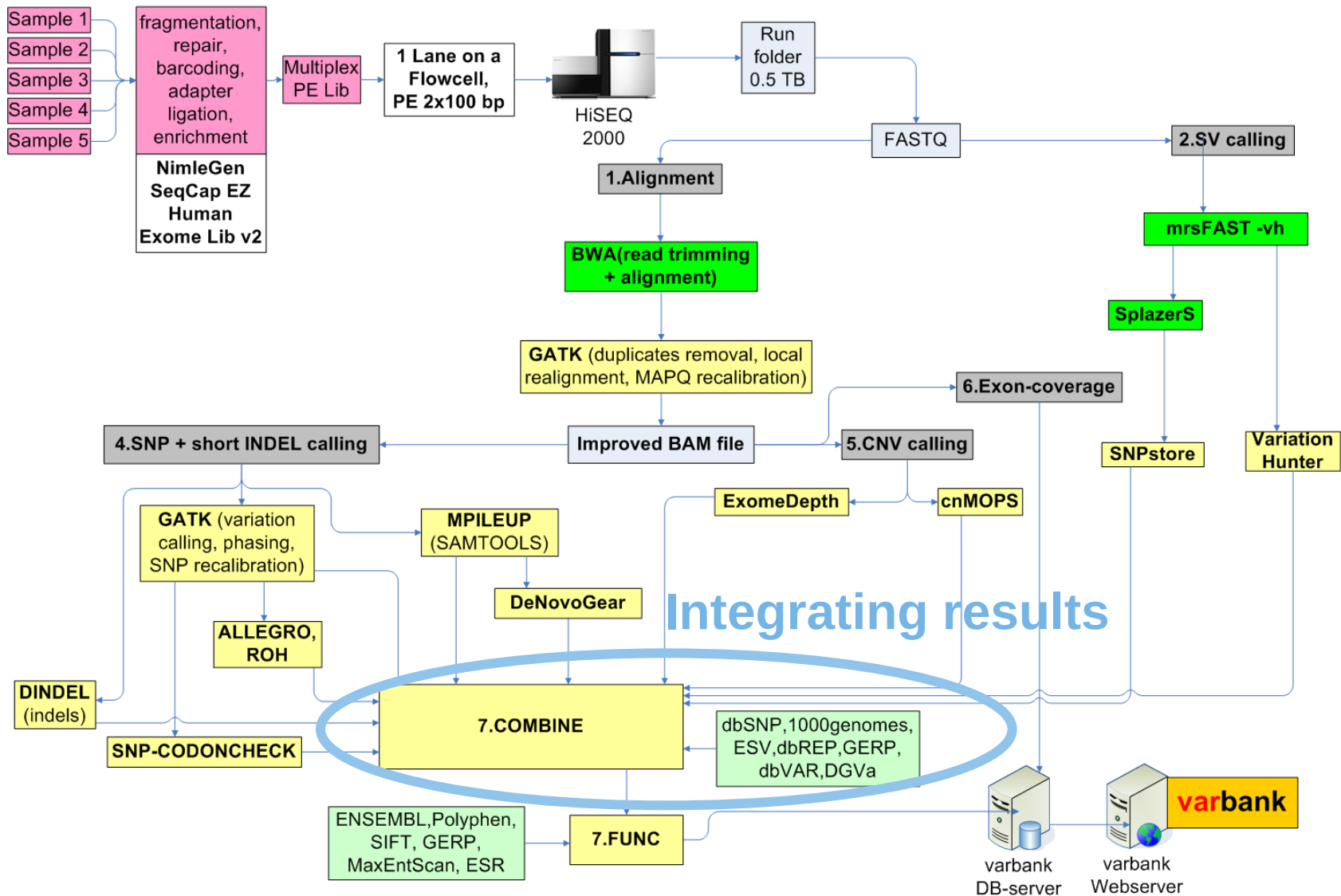
Exome Sequencing Pipeline at the Cologne Center for Genomics



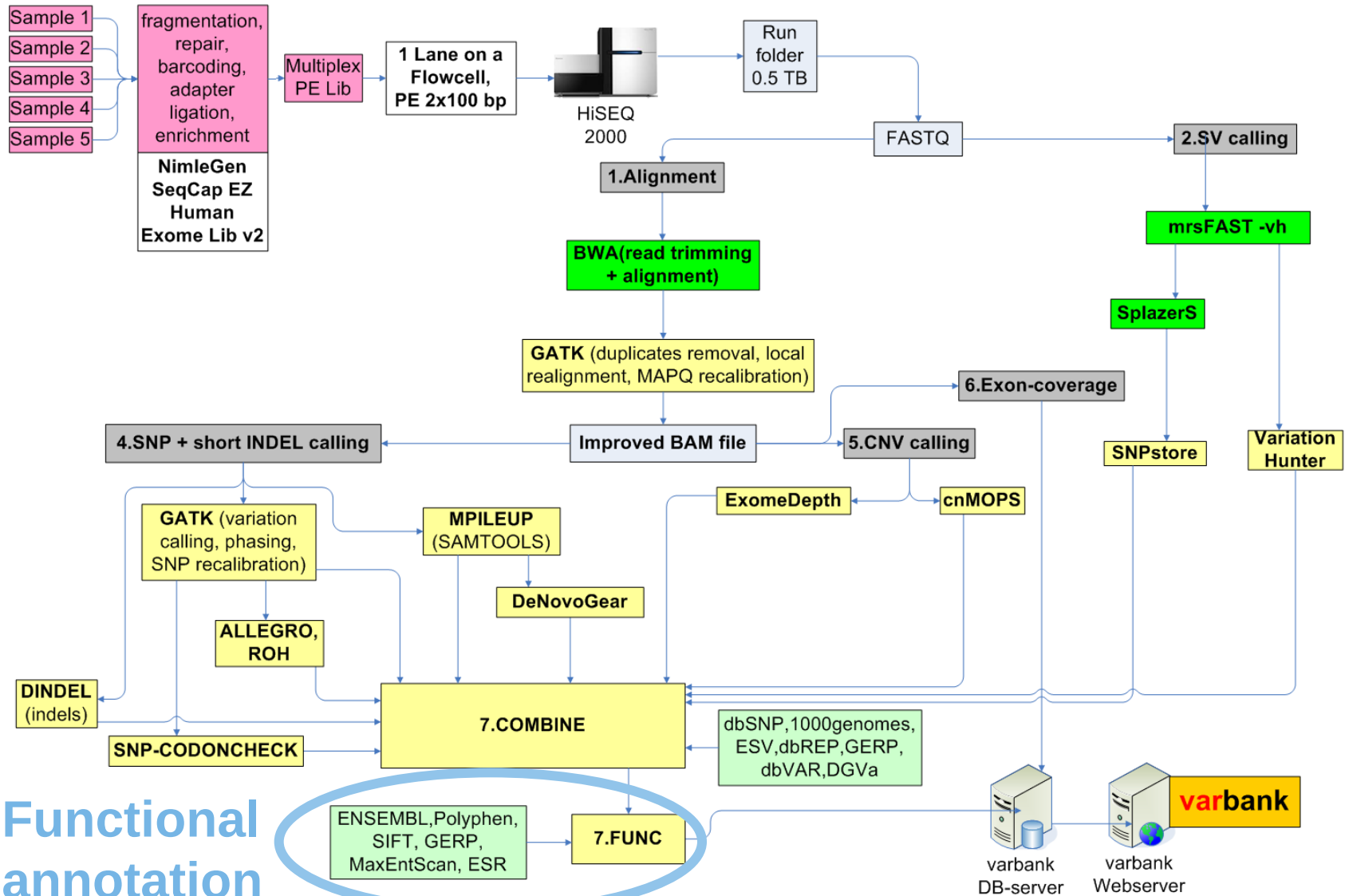
Exome Sequencing Pipeline at the Cologne Center for Genomics



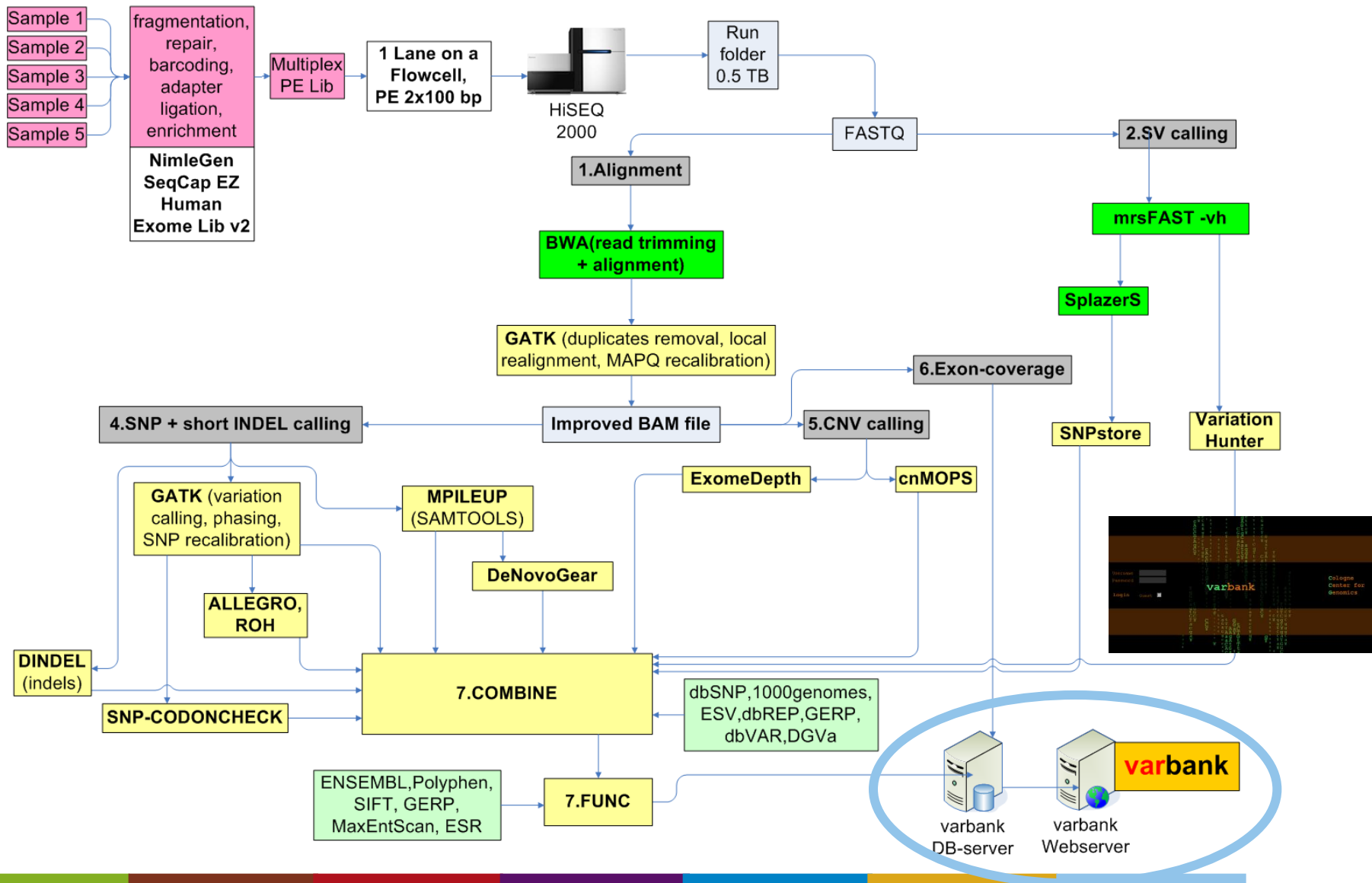
Exome Sequencing Pipeline at the Cologne Center for Genomics



Exome Sequencing Pipeline at the Cologne Center for Genomics



Exome Sequencing Pipeline at the Cologne Center for Genomics



Username

Password

Login

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 maintenance, 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

Select Data Sets														
☐	AB	Prid	Aid	Sid	Enrich	Pipeline	GenBuild	EnsBuild	Pedigree	Sample	Project	Owner	Phenotype	SubPhenTypes
☐														
☐	A	265	1206	5274	Nimbl-SeqCapEZ-V2 37M	2.1_devel	hg19	B68		ROL_0111	RW	Fritz Zimprich	Rolando	2
☐	A	265	1201	5265	Nimbl-SeqCapEZ-V2 37M	2.1_devel	hg19	B68		ROL_0091	RW	Fritz Zimprich	Rolando	2
☐	A	265	1200	5253	Nimbl-SeqCapEZ-V2 37M	2.1_devel	hg19	B68		ROL_0041	RW	Fritz Zimprich	Rolando	2
☐	A	265	1211	5729	Nimbl-SeqCapEZ-V2 37M	2.1	hg19	B68		ROL_0671	RW	Fritz Zimprich	Rolando	2

A

26

☐

A

26

☐

A

24

☐

A

24

1

26

Variation Data												
	EnsProt	Strand	Biotype	HGIC	CCDS	RefSeq	Dist5S5	Dist3S5	RegulationInfo	MutPos	PosFilter	Consequence
17	ENSP00000382526	1	protein_coding	CACNA1C	CCDS53733.1	NM_001167624.1	-89	-15	ESR=20,6,8,13;S3=	CDS.43	CDS	COMPLEX_INDEL;ESE2ESS_CHAN
34	ENSP00000382542	1	protein_coding	CACNA1C		NM_001167625.1	1389	584	BPO=CCGAT,2,3,3,C	INT.43	INTRON	BP_LOSS;CRYPTIC_5SS_ACTIVAT
32	ENSP00000329877	1	protein_coding	CACNA1C	CCDS53736.1	NM_001129830.1	-89	-15	ESR=20,6,8,13;S3=	CDS.43	CDS	COMPLEX_INDEL;ESE2ESS_CHAN
32	ENSP00000353868	-1	protein_coding	PCED1A	CACNA1C					INT.3	INTRON	DEC_3SS_STRONG;ESE_LOSS
37	ENSP00000353077	1	protein_coding	GNRH2						CDS.4	CDS	FRAMESHIFT
17	ENSP00000369705	1	protein_coding	GNRH2						CDS.3	CDS	FRAMESHIFT
33	ENSP00000245983	1	protein_coding	GNRH2						CDS.4	CDS	FRAMESHIFT
16	ENSP00000369704	1	protein_coding	GNRH2						CDS.3	CDS	FRAMESHIFT
30	ENSP00000352003	1	protein_coding	GNRH2						CDS.4	CDS	FRAMESHIFT
17	ENSP00000369705	1	protein_coding	GNRH2						CDS.3	CDS	FRAMESHIFT
37	ENSP00000353077	1	protein_coding	GNRH2						CDS.4	CDS	FRAMESHIFT
30	ENSP00000352003	1	protein_coding	GNRH2						CDS.4	CDS	FRAMESHIFT
33	ENSP00000245983	1	protein_coding	GNRH2						CDS.3	CDS	FRAMESHIFT
16	ENSP00000369704	1	protein_coding	GNRH2						CDS.3	CDS	FRAMESHIFT
72	ENSP00000347184	1	protein_coding	HTT						CDS.1	CDS	FRAMESHIFT
34	ENSP00000448683	1	protein_coding	IL32						CDS.7	CDS	FRAMESHIFT
13	ENSP00000371648	1	protein_coding	IL32						CDS.5	CDS	FRAMESHIFT
30	ENSP00000433747	1	protein_coding	IL32						CDS.6	CDS	FRAMESHIFT
38	ENSP00000324742	1	protein_coding	IL32						CDS.7	CDS	FRAMESHIFT
15	ENSP00000405063	1	protein_coding	IL32						CDS.7	CDS	FRAMESHIFT
34	ENSP00000405364	1	protein_coding	IL32						CDS.6	CDS	FRAMESHIFT
30	ENSP00000008180	1	protein_coding	IL32						CDS.6	CDS	FRAMESHIFT
32	ENSP00000446624	1	protein_coding	IL32	CCDS32379.1	NM_001012636.1		-308	ESR=44,15,44,23;S	CDS.6	CDS	FRAMESHIFT
32	ENSP00000411958	1	protein_coding	IL32	CCDS32377.1	NM_004221.4		-308	ESR=44,15,44,23;S	CDS.7	CDS	FRAMESHIFT

Close

calcium channel, voltage-dependent, L type, alpha 1C subunit

Entrez Gene Summary

This gene encodes an alpha-1 subunit of a voltage-dependent calcium channel. Calcium channels mediate the influx of calcium ions into the cell upon membrane polarization. The alpha-1 subunit consists of 24 transmembrane segments and forms the pore through which ions pass into the cell. The calcium channel consists of a complex of alpha-1, alpha-2/delta, beta, and gamma subunits in a 1:1:1:1 ratio. There are multiple isoforms of each of these proteins, either encoded by different genes or the result of alternative splicing of transcripts. The protein encoded by this gene binds to and is inhibited by dihydropyridine. Alternative splicing results in many transcript variants encoding different proteins. Some of the predicted proteins may not produce functional ion channel subunits. [provided by RefSeq, Oct 2012] [read more](#)

Orphanet rare diseases

[Timothy syndrome](#)[Brugada syndrome](#)

Mouse Genome Database

Mice homozygous for mutations that inactivate the gene do not survive to term. Selective ablation in beta cells resulted in impaired insulin secretion and systemic glucose intolerance. Heterozygotes were hypoactive, showed increased anxiety, and poor motor coordination. [read more](#)

Group A

Allele frequency

15

Variations per gene

1

Genotype

☐ 0/0 ☒ 0/1 ☒ 1/1

ROH spans variation

☐ yes/no

Global Settings

Max Var Frequency

1%

Min Coverage

6

Min Quality

10

Show Transcripts

☒ CCDS/Refseq ☐ Other

Consequences

☒ Protein structure affected ☐ Medium 5'SS/3'SS effect

Region

Excel Export

Page 1 of 15

200

Group A

Allele frequency

15

100

Variations per gene

1

10

Genotype

☐ 0/0

☒ 0/1

☒ 1/1

☒ 1/2

ROH spans variation

☐ yes/no

Group B

Allele frequency

15

100

Variations per gene

1

10

Genotype

☐ 0/0

☒ 0/1

☒ 1/1

☒ 1/2

ROH spans variation

☐ yes/no

Global Settings

Max Var Frequency

1%

MaxTargetDist

100

Min Coverage

6

Transcript Biotype

☒ Protein coding
☐ Other

Min Quality

10

Location

☒ CDS
☐ UTR
☒ INTRON
☒ SPANNING GENES
☐ PROMOTER
☐ INTERGENIC
☒ COMPOSED

Show Transcripts

☒ CCDS/Refseq
☐ Other

Variation Type

☒ SNP
☒ INSERT
☒ CNV
☒ DELETION
☒ INDEL
☒ ROH

Consequences

☒ Protein structure affected
☒ Strong 5'SS/3'SS effect
☒ Cryptic 5'SS/3'SS activation
☐ Medium 5'SS/3'SS effect
☐ ESS/ESE change
☐ Other or no consequences

Region

dbSNP/dbVar

Genes

Exclude Genes

Browse Reads

Dataset:
 Region:
 Transcript:
 Strand:
 Show reads

GAGTTTCTGCTGCTCTCTGTCAGAGCTCTTTCATGTCCTCTTTCCAGTCTACGGAGCCCCAC-GGGGGGACAAGGAGGAGCTGACACCCAGAAAGTGCTCTGAACCCCAATCCTCAAAATGAAGATACTGACACCA



Exon Coverage

Dataset:
 Gene:
 EnsGene:
 Show Exon Coverage

Exon Coverage

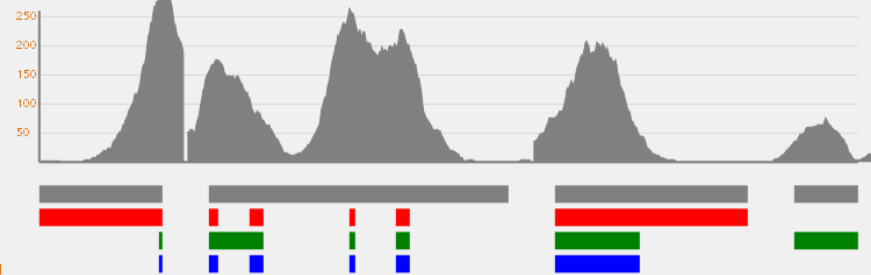
Sample	EnsGene	Gene	Transcript	Exon	Chr	Start	Stop	Min_Cov	Av_Cov	Max_Cov	Perc_Cov
ROL_0091	ENSG000000008517	IL32	ENST000000008180	1	16	3115632	3115688	42	63.51	94	100
ROL_0091	ENSG000000008517	IL32	ENST000000008180	2	16	3115785	3115827	240	274.86	297	100
ROL_0091	ENSG000000008517	IL32	ENST000000008180	3	16	3117378	3117416	0	161.49	176	97.44
ROL_0091	ENSG000000008517	IL32	ENST000000008180	4	16	3117985	3118011	240	254.3	263	100
ROL_0091	ENSG000000008517	IL32	ENST000000008180	5	16	3118181	3118240	194	211.83	230	100
ROL_0091	ENSG000000008517	IL32	ENST000000008180	6	16	3118991	3119820	0	66.06	208	88.92
ROL_0091	ENSG000000008517	IL32	ENST00000325568	1	16	3115298	3115495	0	1.27	3	73.23
ROL_0091	ENSG000000008517	IL32	ENST00000325568	2	16	3115785	3115827	240	274.86	297	100
ROL_0091	ENSG000000008517	IL32	ENST00000325568	3	16	3117378	3117416	0	161.49	176	97.44
ROL_0091	ENSG000000008517	IL32	ENST00000325568	4	16	3117555	3117614	72	86	111	100
ROL_0091	ENSG000000008517	IL32	ENST00000325568	5	16	3117985	3118011	240	254.3	263	100
ROL_0091	ENSG000000008517	IL32	ENST00000325568	6	16	3118991	3119820	0	66.06	208	88.92
ROL_0091	ENSG000000008517	IL32	ENST000000008180	1	16	3115632	3115688	42	63.51	94	100
ROL_0091	ENSG000000008517	IL32	ENST000000008180	2	16	3115785	3115827	240	274.86	297	100
ROL_0091	ENSG000000008517	IL32	ENST000000008180	3	16	3117378	3117416	0	161.49	176	97.44
ROL_0091	ENSG000000008517	IL32	ENST000000008180	4	16	3117985	3118011	240	254.3	263	100
ROL_0091	ENSG000000008517	IL32	ENST000000008180	5	16	3118181	3118240	194	211.83	230	100
ROL_0091	ENSG000000008517	IL32	ENST000000008180	6	16	3118991	3119820	0	66.06	208	88.92
ROL_0091	ENSG000000008517	IL32	ENST00000325568	1	16	3115298	3115495	0	1.27	3	73.23
ROL_0091	ENSG000000008517	IL32	ENST00000325568	2	16	3115785	3115827	240	274.86	297	100
ROL_0091	ENSG000000008517	IL32	ENST00000325568	3	16	3117378	3117416	0	161.49	176	97.44
ROL_0091	ENSG000000008517	IL32	ENST00000325568	4	16	3117555	3117614	72	86	111	100
ROL_0091	ENSG000000008517	IL32	ENST00000325568	5	16	3117985	3118011	240	254.3	263	100
ROL_0091	ENSG000000008517	IL32	ENST00000325568	6	16	3118991	3119820	0	66.06	208	88.92

View 1 - 197 of 197

Gene coverage

Dataset:
 Gene:
 X-scaling:
 Y-scaling:
 Show coverage

ROL_0091
 ENSG000000008517 (1)



Username

Password

Login

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

varbank

Cologne
Center for
Genomics