



NGS-based HLA and Bloodgroup Typing - One-Stop Shop (?)

Prof. Andre Franke
Institute of Clinical Molecular Biology | Kiel University

Swisstransfusion, 25.8.16 in Bern





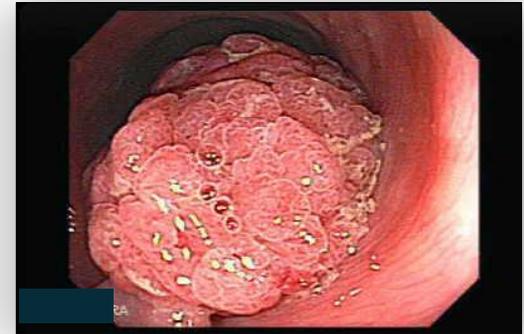
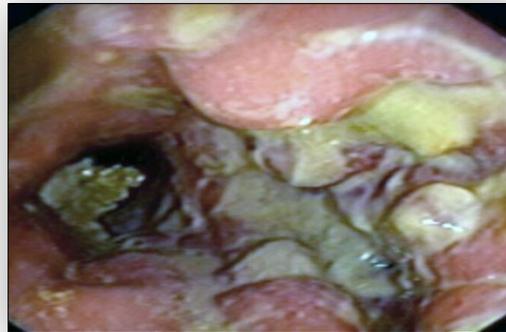
Focus on two methods

- 1) Imputation of genotypes
- 2) In-solution capture of DNA using RNA baits



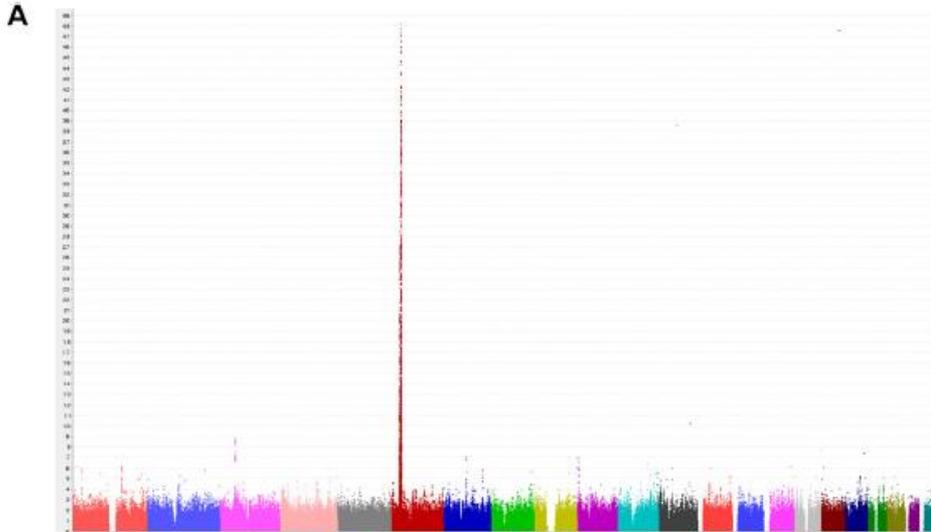
Inflammation

An important mechanism





Importance of HLA association



PSC (Ulcerative colitis)



disease

diabetes

sclerosis

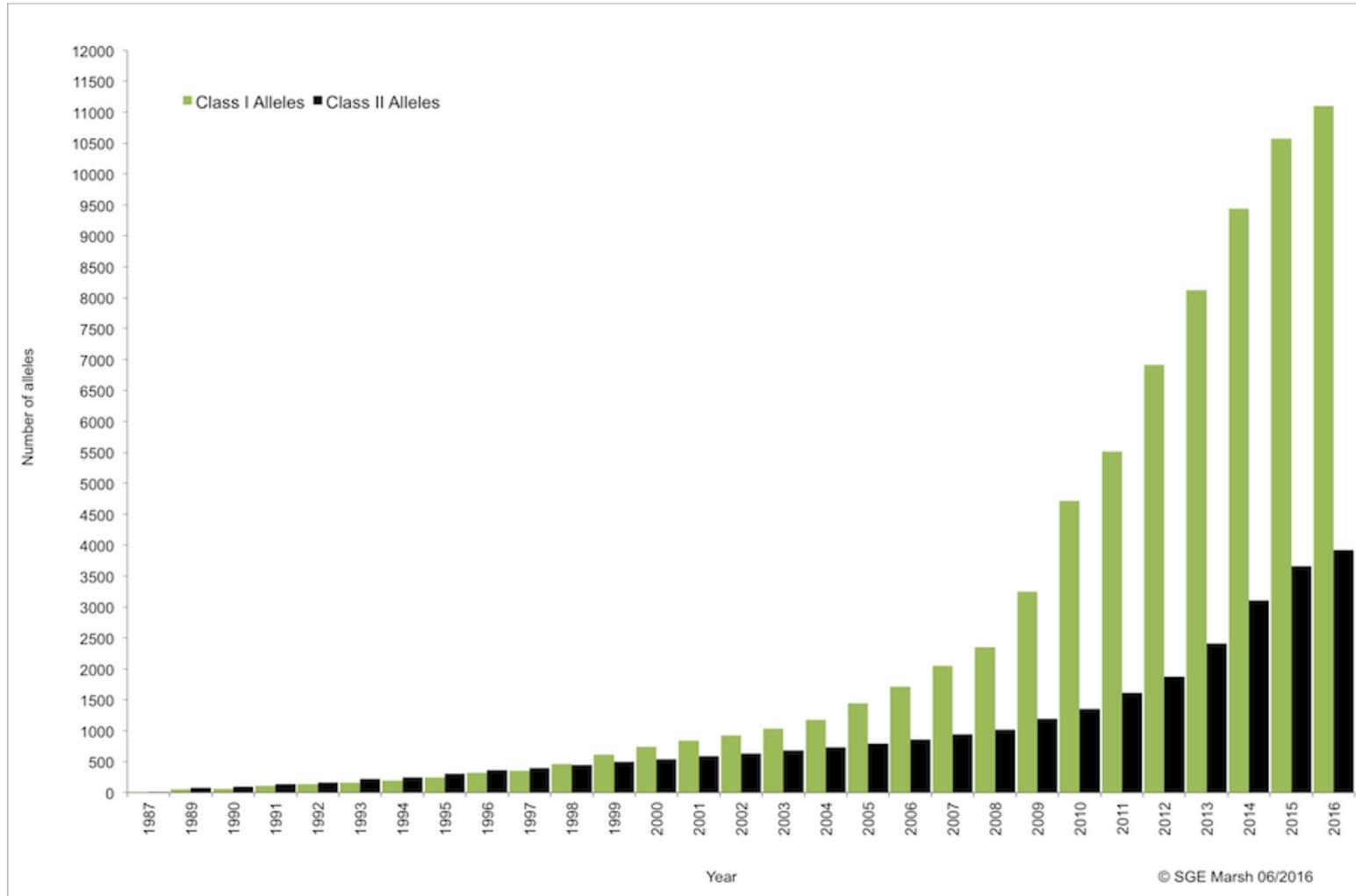
Number of different alleles

- ▶ Number of alleles annotated for the most important HLA loci, IMGT/HLA database v3.25 , July 2016

HLA Class I					
<i>Gene</i>	<i>A</i>	<i>B</i>	<i>C</i>		
Alleles	3,492	4,358	3,111		
Proteins	2,480	3,221	2,196		
HLA Class II					
<i>Gene</i>	<i>DRB</i>	<i>DQA1</i>	<i>DQB1</i>	<i>DPA1</i>	<i>DPB1</i>
Alleles	2,135	73	940	43	671
Proteins	1,569	33	647	21	552



Growing number of alleles





„The HLA is a monster“*



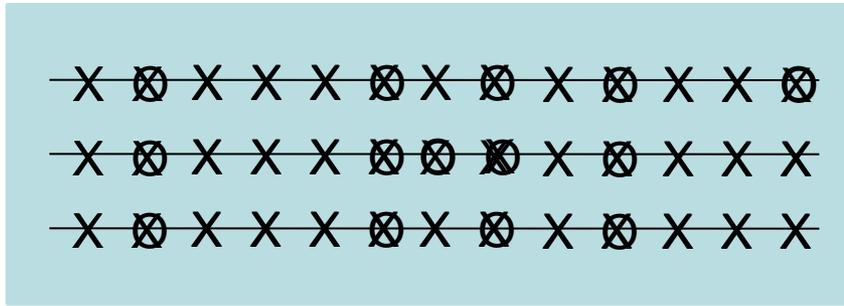
*Prof. Tom H. Karlsen 2008, Oslo, Norway

Genotype Imputation

Reference



Custom



- O Missing genotypes
- X Available genotype data

Human Mutation, 2008

A Comprehensive Evaluation of SNP Genotype Imputation

Michael Nothnagel^{1,2}, David Elinghaus^{1,2}, Stefan Schreiber^{1,2}, Michael Krawczak^{1,2}, Andre Franke^{1,2*}
¹ Institute of Medical Informatics and Statistics, ² Institute of Clinical Molecular Biology, both Charité-Universitätsmedizin Berlin, Germany
 * equal contribution, ** equal senior authorship

Introduction.

SNP imputation might reduce costs in exploratory association studies and facilitate the detection of new causative variants. A comprehensive validation and comparison of available programs is lacking. We assessed systematically and at a genome-wide level how imputation performed in our own data.

Samples.

DNA samples of 241 male and 218 female unrelated blood donors (459 in total) were obtained from the HapMap biobank (1). The individuals, their parents and grandparents were all born in Germany.

Genotyping.

All samples were genotyped with each of the following three arrays, adding complementary data:

- Affymetrix SNP Array 5.0 (500k)
- Affymetrix SNP Array 6.0 (1.000k)
- Illumina Human1M-D3 bead array (550k)

 We restricted our analysis to autosomal markers.

Quality control.

- Call rate >95% per individual, >95% per marker
- Minor allele frequency >0.01
- Hardy-Weinberg equilibrium >= 0.01
- Average identity-by-state between samples within the three-fold array-wide IQR

Marker Sets and Imputation Reference

Marker sets used for imputation benchmarking.

We split number of autosomal, QC-passed SNPs (total) in each array and the percentage included in the study in parenthesis.

Upper right half: number of overlapping SNPs between any two arrays (average genotype concordance rate in parenthesis).

Lower left half: number of SNPs that were unique to an array type and had phasing information available in HapMap CEU (imputable SNPs); these SNPs were used for benchmarking.

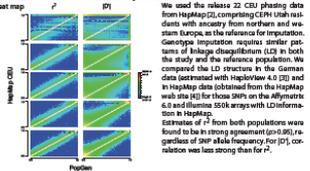
Top-left: imputable SNPs unique to columns array type. Bottom-left: imputable SNPs unique to row array type. (Total number of unique SNPs and the percentage included in the study in parenthesis).

Array Type	No. SNPs	Array Type	No. SNPs	Array Type	No. SNPs
Affymetrix 5.0	500,000 (94.1% of 532,485)	Affymetrix 6.0	1,000,000 (94.1% of 1,063,700)	Illumina 1M-D3	550,000 (94.1% of 584,500)
Affymetrix 5.0 & 6.0	1,499,999 (94.1% of 1,596,185)	Affymetrix 5.0 & 1M-D3	1,049,999 (94.1% of 1,116,485)	Affymetrix 6.0 & 1M-D3	1,549,999 (94.1% of 1,645,200)
Affymetrix 5.0, 6.0 & 1M-D3	2,049,997 (94.1% of 2,178,670)				

Imputation reference and its representativeness.

We used the reference CEU phasing data from HapMap (2), comprising CEU Utah residents with ancestry from northern and western Europe, as the reference for imputation. Genotype imputation requires similar patterns of linkage disequilibrium (LD) in both the study and the reference population. We compared the LD structure in the German data (asstrued with HaploView 4.0 (3)) and in HapMap data (obtained from the HapMap web site (4)) for those SNPs on the Affymetrix 6.0 and Illumina 550k arrays with LD information in HapMap.

Estimates of r^2 from both populations were found to be in strong agreement (0.95-1.0), regardless of SNP allele frequency; for r^2 correlation was less strong than for r^1 .

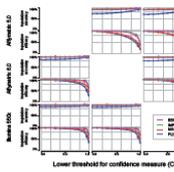


Imputation Benchmarking

Effect of varying CT values.

Dependence of imputation accuracy and efficacy on the confidence threshold.

Rows: Imputation basis; Columns: Imputation target



Programs and performance measures.

Programs: BEAGLE v1.0.10, IMPUTE v1.0.2, MACH v1.0.102, PLINK v1.0.02

Confidence: r^2 based (CEU), program-specific confidence measure for imputation

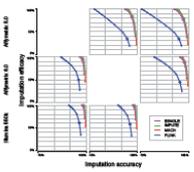
Imputation efficacy: proportion of imputable SNPs with confidence exceeding given CT

Imputation accuracy: concordance rate between the imputed and observed genotypes

Accuracy-efficacy trade-off.

Simultaneous measurement of imputation accuracy and efficacy with varying confidence thresholds.

Rows: Imputation basis; Columns: Imputation target



Imputation accuracy and efficacy at default CT values.

Left: Columns: Imputation basis and number of QC-SNPs with HapMap phasing data available. For each combination of Imputation basis (row) and target (column).

Top: Imputable SNPs; Center: Imputation efficacy; Bottom: Imputation accuracy

Default CT values: BEAGLE: 0.90, IMPUTE: 0.90, MACH: 0.30, PLINK: 0.90

Imputation basis \ Imputation target	Imputation basis				Imputation target			
	CEU	GER	FIN	YRI	CEU	GER	FIN	YRI
BEAGLE	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
IMPUTE	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
MACH	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
PLINK	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94

Conclusions

Ha-Map-based imputation. HapMap CEU-based imputation can reliably infer missing genotypes in a population of Northern European descent. MACH and IMPUTE turned out to be the programs of choice to warrant both high accuracy and high efficacy at only a minor loss of accuracy. BEAGLE may also be a good choice for some applications. Nothing can be said, however, about populations which HapMap CEU only partially represents, e.g. samples from the South and East of Europe (5) and about the performance of imputation in regions devoid of HapMap markers.

Use of imputed genotypes.

Imputed genotypes are not actually observed genotypes. The ambiguity of their prediction has to be included explicitly in their interpretation. One way to address this issue in the context of significance testing would be to use the posterior genotype probability instead of discrete genotypes. For example, in linear or logistic regression models, the need to include ambiguity applies particularly in the context of whole-haplotype association. Simply using marker-wise best guesses of genotypes as a makeshift strategy.

Acknowledgements / References

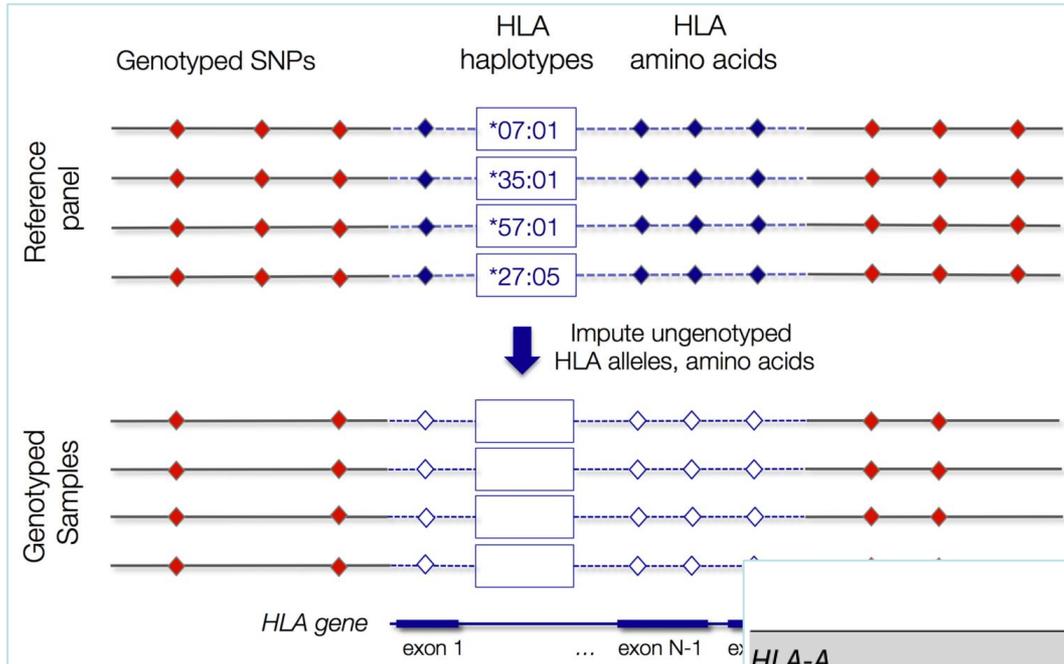
This work was supported by the German National Genome Research Network (NGFN).

- 1) Kruczkia M et al (2006) Community Genet 9:55-61
- 2) Istrail S et al (2007) Nature 449:95-101
- 3) Barrett JC et al (2008) Bioinformatics 23:1363-5
- 4) The International HapMap Consortium (2005) Nature 437:129-320
- 5) Browning SB, Browning BL (2007) Am J Hum Genet 81:1084-97
- 6) Marchini J et al (2007) Nat Genet 39:906-13
- 7) Li J, Abecasis GR (2006) Am J Hum Genet 79:230-5
- 8) Purcell S et al (2007) Am J Hum Genet 81:558-95
- 9) Luo Y et al (2008) Eur J Hum Genet 16:1242-9
- 10) Development Core Team (2008) Foundation, Vienna, Austria

Please see also: Nothnagel et al (2008) A Comprehensive Evaluation of SNP Genotype Imputation. Hum Genet, in review.

Web: <http://capella.uni-kiel.de/> E-mail: nothnagel@medinf.uni-kiel.de

SNP2HLA



	CEU/CEPH	YRI	CHB+JPT
HLA-A	99.1%	69.9%	98.1%
HLA-B	96.8%	90.5%	65.6%
HLA-C	99.1%	98.4%	68.8%
HLA-DQA1	98.5%	64.9%	96.3%
HLA-DQB1	99.1%	96.1%	96.5%
HLA-DRB1	96.9%	20.3%	92.3%
All loci	98.3%	72.9%	86.4%

HLA fine mapping

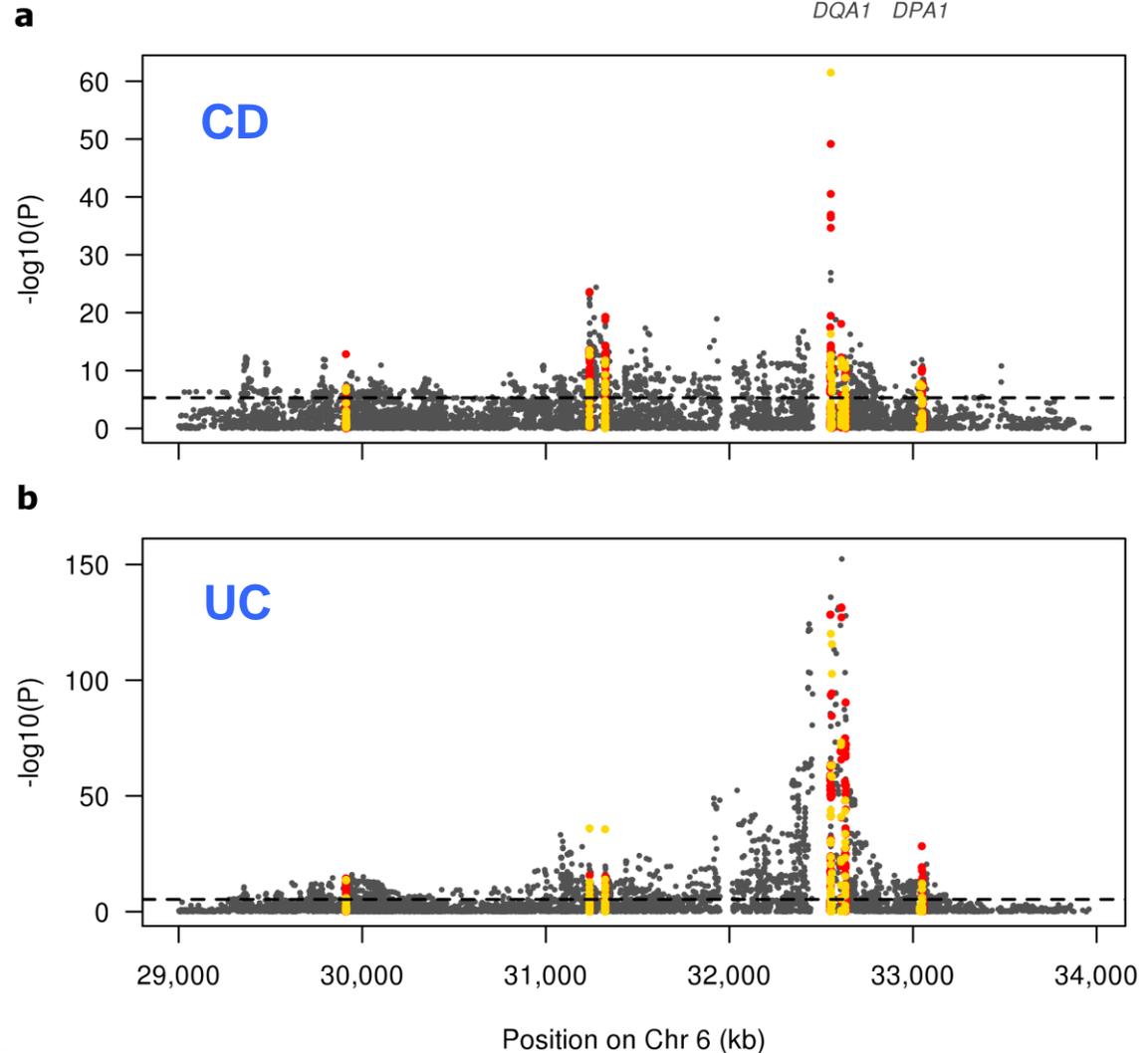


34,241 controls

18,405 CD

14,308 UC

- 8939 SNPs
- 741 AA
- 90/138 class. genes





Classical analysis

IKMB Sequencing Center (since 2004)

Next Generation Sequencing Unit @ICMB

3x Illumina HiSeq2500
1x Illumina HiSeq3000
1x Illumina HiSeq4000



1x Illumina NextSeq



2x Illumina MiSeq



Raw NGS Data



300 TB

Storage @ICMB

RAW/Project Data



900 TB

High Throughput Compute (HTC) Cluster @RZ CAU

Headnode



40 TB XFS

EMC Isilon X200/NL400



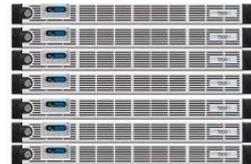
600TB OneFS

NFS - Storage

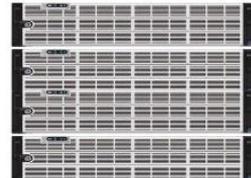


218 TB XFS

32 compute nodes
8 cores (total: 256)
32 Gb RAM
1,7 Tb scratch per node



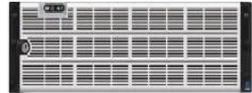
42 compute nodes
16 cores (total: 672)
128 Gb RAM
3 Tb scratch per node



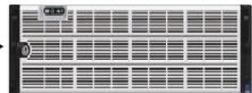
BWA, GATK,
miRDeep2,
R/Bioconductor

Oracle SAM-QFS @RZ CAU

Hierarchical Storage Management



Oracle Storage FS1 @RZ



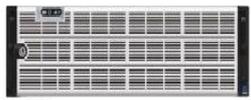
Tape library cache server (100TB)

Oracle StorageTek SL8500 @RZ



expandable to 30 PB

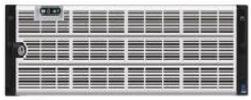
High RAM compute @ICMB



80 cores
1.5 Tb RAM
32 TB

Spades
velvet
Prokka
PhyloPhlan
Harvest

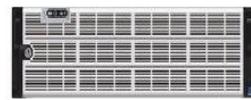
Virtualization platform @ICMB



80 cores
756 Gb RAM
144 TB

OpenStack
KVM

Bioinformatics Webservers @RZ



32 cores
512 Gb RAM
86 TB

GNOS Annai
Grabblur
HGMD
PolyPhen2
OwnCloud

FPGA - parallel computing @RZ

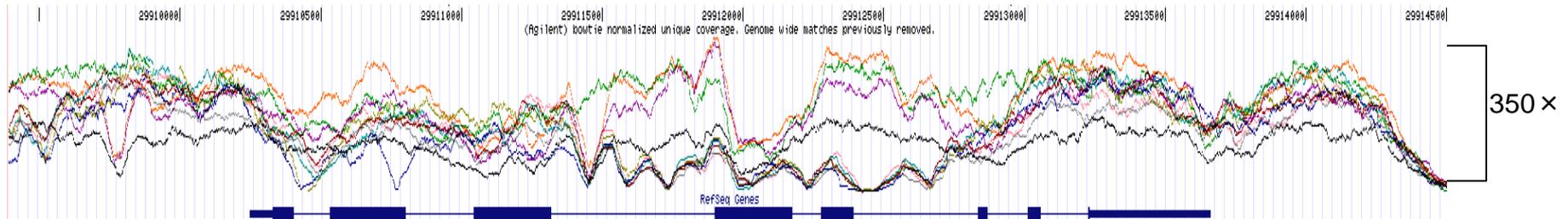
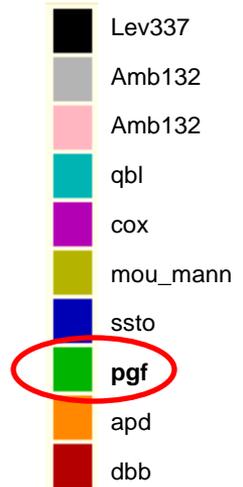


128 FPGAs
2x 256 Gb RAM each

Swa
BLAST
ShapelT

2012 study with reference-based design

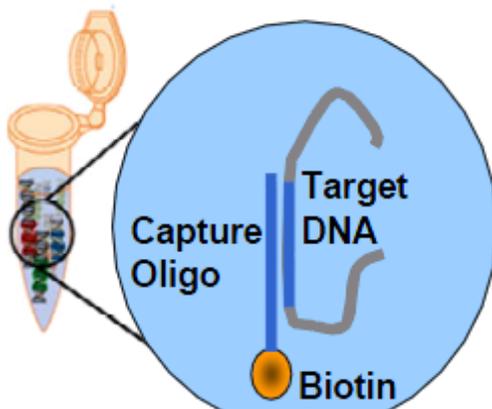
- ▶ Alignment of the enriched NGS runs to pgf haplotype
- ▶ Amb132, a heterozygous sample (2 replicates)
- ▶ Lev337, a heterozygous sample
- ▶ 7 haploid cell line samples of known haplotype
Horton et al. Immunogenetics (2008) 60: 1-18



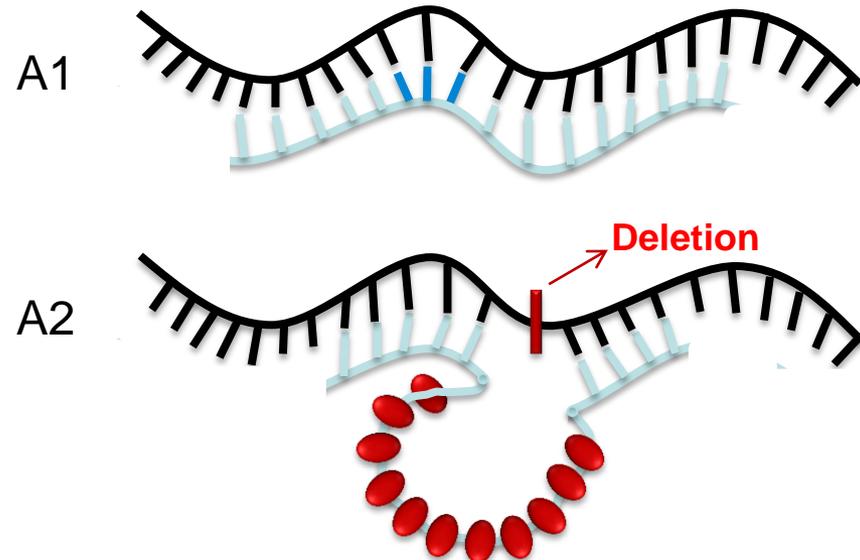


Setup of benchmark

- ▶ 357 samples (enriched for diversity)
- ▶ 48 samples/lane on HiSeq 2500
- ▶ SureSelect Automated LibraryPrep and Capture System



In-solution



Diversity of test samples

	Sample #	A	B	C	DPA	DPB	DQA	DQB	DRB
NGS Kiel	299	74	111	53	73	20	17	5	42
Classical Typing									
Sanger	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
* Fluidigm	192	N.A							
** Roche 454									
Amplicon Seq.	173	29	63	26	-	22	-	19	19
*** PNAS Amplicon									
Seq.	79	41	88	35					41

* Moonsamy PV *et al.* 2013. High throughput HLA genotyping using 454 sequencing and the Fluidigm Access Array™ System for simplified amplicon library preparation. *Tissue Antigens*. 81(3):141-9.

** Martin Danzer *et al.* 2013. Rapid, scalable and highly automated HLA genotyping using next-generation sequencing: a transition from research to diagnostics. *BMC Genomics* 2013, 14:221

*** Chunlin Wang *et al.* 2012. High-throughput, high-fidelity HLA genotyping with deep sequencing. *PNAS* 109(22):8676-81

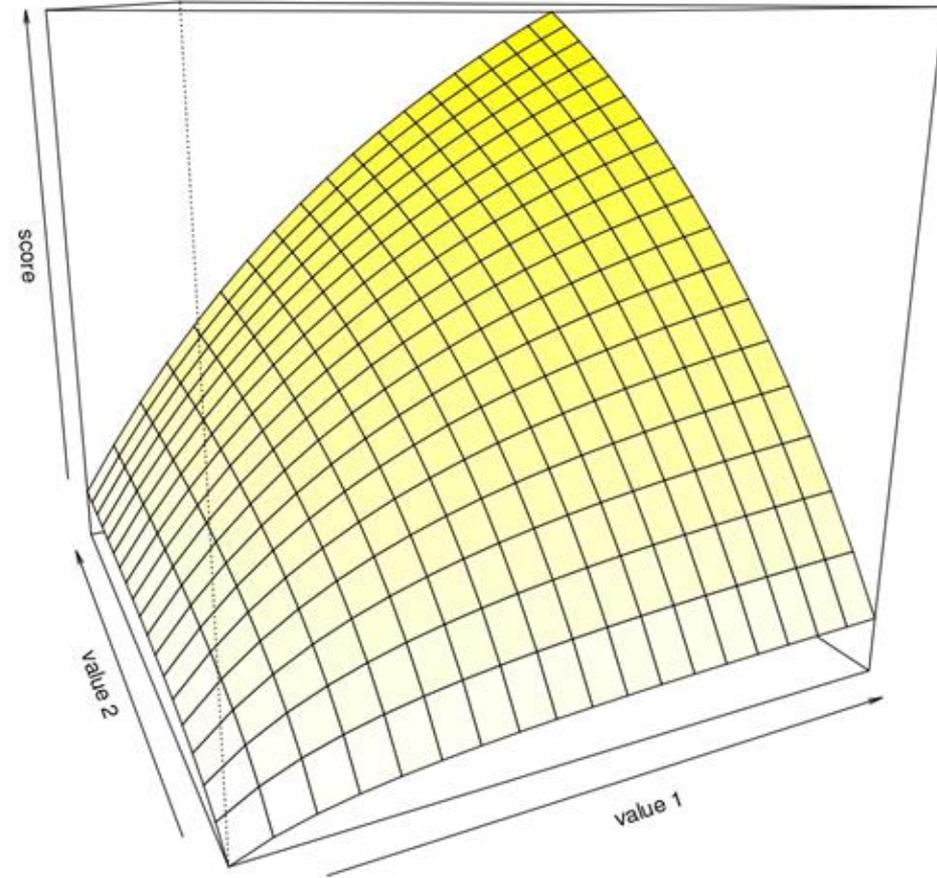
HLA allele calling

P: parameters characterizing the mapping

w: weighting of the different parameters

$P = \{ asm, req, msl, mppr, auc \}$

$w = \{ 0.5, 1, 0.1, 0.1, 1 \}$



$$H = \frac{\left(\sum_{k=1}^5 \frac{1}{w_k} \right) * \prod_{k=1}^5 P_k}{\left(\sum_{k=1}^5 \frac{1}{w_k} P_k \right)}$$

$$H = \frac{(2 + 1 + 10 + 10 + 1) * asm * req * msl * mppr * auc}{2 * asm + 1 * req + 10 * msl + 10 * mppr + 1 * auc}$$



HLA calling success rate

Table 1 Benchmark statistics for allele calls that could be achieved in a fully automated manner at three-field resolution (note that some of the reference samples only had two-field resolution, Supplementary Table S2)

Locus	Genotypes available	Different alleles	Allele call rate		Mean alternatives/sample
			Exact	Including alternatives	
A	341	114	0.98	0.99	2.21
B	344	122	0.99	1.00	1.03
C	285	53	0.98	0.99	1.58
DRB1	307	80	0.98	0.98	1.08
DQA1	152	20	1.00	1.00	0.06
DQB1	264	18	0.99	0.99	0.06
DPA1	87	5	0.98	0.98	0.00
DPB1	209	40	0.97	0.97	0.57

Columns - *Locus*: Lists the typed classical gene locus. *Genotypes available*: Shows the number of samples for which reference genotypes were available. Only alleles for which a reference call was available were used for calculating the call rate. *Different alleles*: Lists the number of different alleles within the reference data set for each locus. The automated calling generates two probable allele calls per diploid sample. *Allele calling exact*: Shows the validated allele call rate that was achieved by the automated calling. *Allele calling including alternatives*: Shows the call rate when including possible alternative alleles (ambiguities). *Mean alternatives/sample*: Shows the average number of alternative allele calls (ambiguities) per sample.

Evolution of HLAssign

```

mwittig@MW-X1C:~/ramDisk
mwittig@MW-X1C:~/ramDisk$ cat dummy.txt+
for i in *.fastq.gz
do
zcat $i | awk '{if( (NR+3)%4==0 ) {split($0, a, " "); print">"a[1]"/"NR if( (NR+2)%4==0 ) {print(substr($0,1,100)
}}}' > ${TMPDIR}/reads_to_map.fasta
done

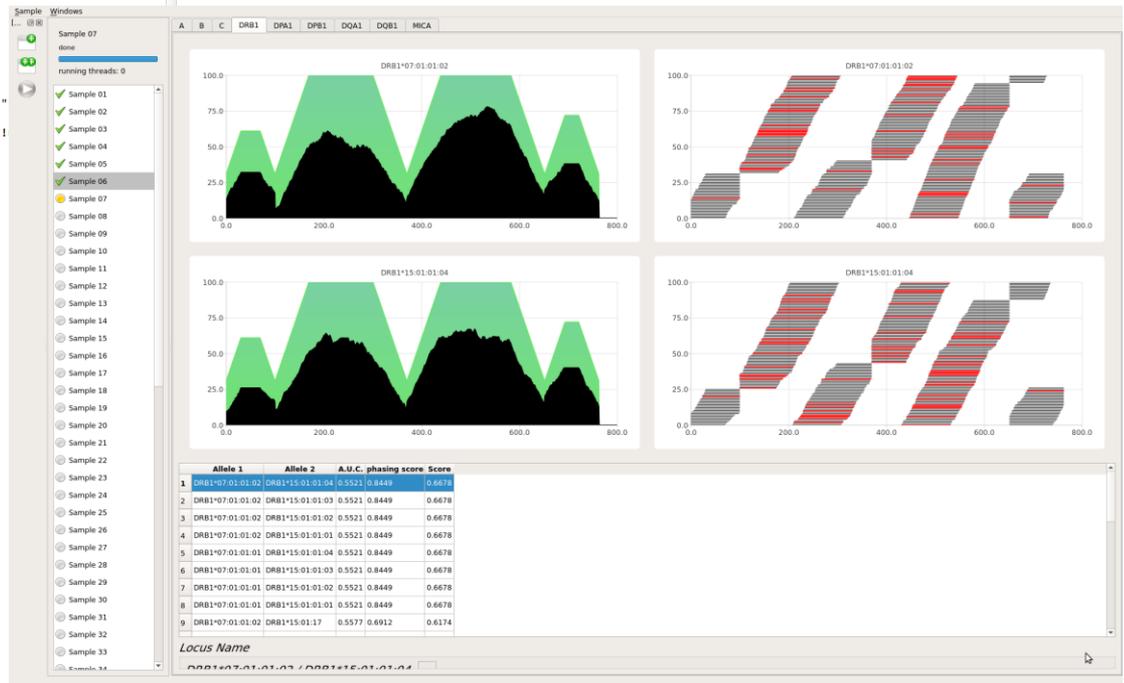
# create required directories
mkdir -p ${THERESULTDIR}/pics
cd ${TMPDIR}
mkdir ${TMPDIR}/pics
mkdir ${TMPDIR}/scratch

# run the analysis
nextcallhla --cdna ${THEALLELEDIR}/THELOCUS_nuc.txt --gdna ${THEALLELEDIR}/THELOCUS_gen.txt --scratch ${TMPDIR}/scr
atch --out ${TMPDIR}/result.txt --outPic ${TMPDIR}/pics --bias-alleles /foo/bar/incomplete_alleles.txt --amb-table
/foo/bar/ambiguity_table.txt --reads ${TMPDIR}/reads_to_map.fasta --read-length 100

cp ${TMPDIR}/pics/* ${THERESULTDIR}/pics/
cp ${TMPDIR}/result.txt ${THERESULTDIR}/THELOCUS_calls_unsorted.txt

# extract calls and store in a file called the_calls.txt
for j in A B C DRB1 DPA1 DPB1 DQA1 DOB1 MICA
do
THECALL='grep -A 1 "[Result]" ${THERESULTDIR}/${j}_calls_unsorted.txt | tail -1'
echo $i $j $THECALL | awk 'BEGIN {IFS=" "}; END {if( NF <= 7 ){print $0} else {print $1" "
}}' >> the_calls.txt
for k in `sort -k 8,8n ${THERESULTDIR}/${j}_calls_unsorted.txt | awk '{if( NF == 8 && $8 !
d -S | cut -f 1 | sed -re 's/"/-/g''
do

```



Graphical User Interface (GUI)





Prices

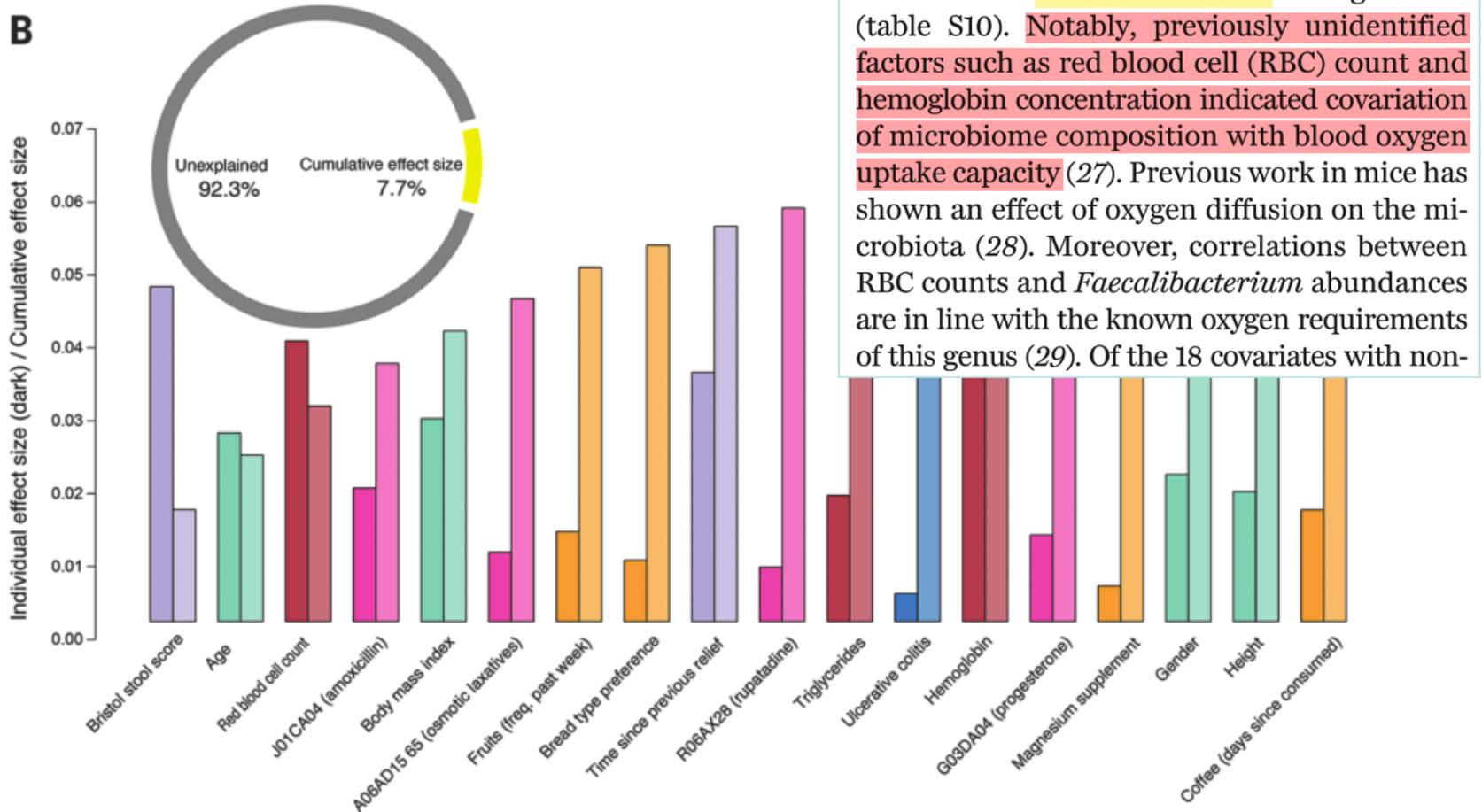
- Luminex – 100 € per test, needs Luminex-machine, throughput?
- Sanger-based technologies - €100-500 per test
- Service-only competition: DKMS (50 € per test), Univ. Munich (50 € per test); Histogenetics LLC
- <http://www.illumina.com/clinical/hla-sequencing.html>
List price 300 € (long-range PCR)
- <http://www.gendx.com>
List price 120 €/sample, up to 384 samples on MiSeq
Hi Throughput: HiSeq 2500 oder NextSeq, 2 x 384 samples
- <http://www.omixon.com/holotype-hla/>
Software from Omixon, Kit by CHOP, 192/240 samples per MiSeq run (7/5 loci)
Hi Throughput price (5000 samples): ca. 55 €/sample regardless whether 5 or 7 loci

Each of these tests have their own problems apart from the price. Problems that can lead to fatal errors. (30%-50% of stem cell transplant recipients die from graft-versus-host disease!)

What about blood group antigens?



Blood groups vs. Microbiome?



Poor reference databases...

Table 1. *MNS* alleles with single nucleotide polymorphisms that generate blood group antigens.

A. *GYP A*: Reference allele *MNS*01* encodes M, En^a, ENEH, ENEP, ENAV, ENDA, ENEV.

Note: In most cases, the nucleotide changes also can occur on an N allele; these nucleotide changes are not given.

Phenotype	Allele name	Nucleotide change	Intron/ Exon	Amino acid change	Comments
MNS:1 or M+	<i>GYP A*01</i> <i>GYP A*M</i>	59C; 71G; 72T	2	Ser20, Gly24	
MNS:2 N+	<i>GYP A*02</i> <i>GYP A*N</i>	59C>T; 71G>A; 72T>G	2	Ser20Leu, Gly24Glu	
MNS:1,-2,8† M ^c +	<i>GYP A*08</i> <i>GYP.Mc</i>	71G>A,72T>G	2	Gly24Glu	
MNS:7,9,-40 Vw+	<i>GYP A*09</i> <i>GYP A*Vw</i>	140C>T	3	Thr47Met	
MNS:-1,-2,11 M ^g +	<i>GYP A*11</i> <i>GYP A*Mg</i>	68C>A	2	Thr23Asn	
MNS:12 Vr+	<i>GYP A*12</i> <i>GYP A*Vr</i>	197C>A	3	Ser66Tyr	
MNS:14 Mt(a+)	<i>GYP A*14</i> <i>GYP A*Mta</i>	230C>T	3	Thr77Ile	
MNS:16 Ri(a+);	<i>GYP A*16</i> <i>GYP A*Ria</i>	226G>A	3	Glu76Lys	
MNS:18 Ny(a+)	<i>GYP A*18</i>	138T>A	3	Gln46Glu	
MNS:7,19,-40 Hut+	<i>GYP A*19</i> <i>GYP A*Hut</i>	140C>A	3	Thr47Lys	



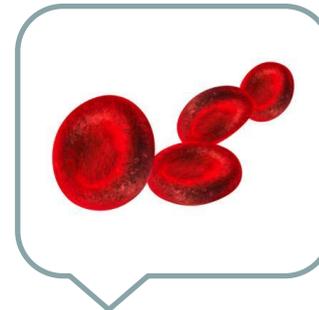
Imputation benchmark

Coa/Cob	5.4%	Jka/Jkb	48.3%	RhCc	31.8%
Cw	3.5%	KEL Kk	3.8%	RhD	18.2%
Dia/Dib	0.1%	KEL Kpa/Kpb	1.0%	RhEe	13.9%
Doa/Dob	37.5%	KEL Jsa/Jsb	0.0%	Sc1/Sc2	0.0%
Fya/Fyb	44.1%	Lua/Lub	3.2%	Vel	0.1%
HPA1a/HPA1b	16.5%	LWa/LWb	1.0%	Wra/Wrb	0.0%
HPA5a/HPA5b	8.3%	MNS MN	45.4%	Yta/Ytb	4.6%
Ina/Inb	0.0%	MNS Ss	31.9%		



NGS experiment

Chromosome	Gene Name	Blood Group
9	<i>ABO</i>	ABO blood group
7	<i>AQP1</i>	Colton blood group
17	<i>SLC4A1</i>	Diego blood group
12	<i>ART4</i>	Dombrock blood group
1	<i>CD55</i>	Duffy blood group
1	<i>ACKR1</i>	Duffy blood group
19	<i>FUT1</i>	galactoside 2-alpha-L-fucosyltransferase, H blood group
2	<i>GYPC</i>	Gerbich blood group
9	<i>AQP3</i>	Gill blood group
X	<i>GATA1</i>	globin transcription factor 1
3	<i>B3GALNT1</i>	globoside blood group
6	<i>GCNT2</i>	I blood group
11	<i>CD44</i>	Indian blood group
15	<i>SEMA7A</i>	John Milton Hagen blood group
4	<i>ABCG2</i>	Junior blood group
7	<i>KEL</i>	Kell blood group
18	<i>SLC14A1</i>	Kidd blood group
1	<i>CR1</i>	Knops blood group
19	<i>KLF1</i>	Kruppel-like factor 1
19	<i>ICAM4</i>	Landsteiner-Wiener blood group
2	<i>ABCB6</i>	Langereis blood group
19	<i>BCAM</i>	Lutheran blood group
4	<i>GYPA</i>	MNS blood group
4	<i>GYPB</i>	MNS blood group
4	<i>GYPE</i>	MNS blood group
19	<i>BSG</i>	Ok blood group
22	<i>A4GALT</i>	Pk antigen of blood histogroup P
11	<i>CD151</i>	Raph blood group
1	<i>RHCE</i>	Rh blood group, CcEe antigens
1	<i>RHCE</i>	Rh blood group, CcEe antigens
1	<i>RHCE</i>	Rh blood group, CcEe antigens
1	<i>RHD</i>	Rh blood group, D antigen
1	<i>RHD</i>	Rh blood group, D antigen
6	<i>RHAG</i>	Rh-associated glycoprotein, ammonium transporter Rh type A
1	<i>ERMAP</i>	Scianna blood group
X	<i>XK</i>	X-linked Kx blood group
X	<i>XG</i>	Xg blood group
7	<i>ACHE</i>	Yt blood group

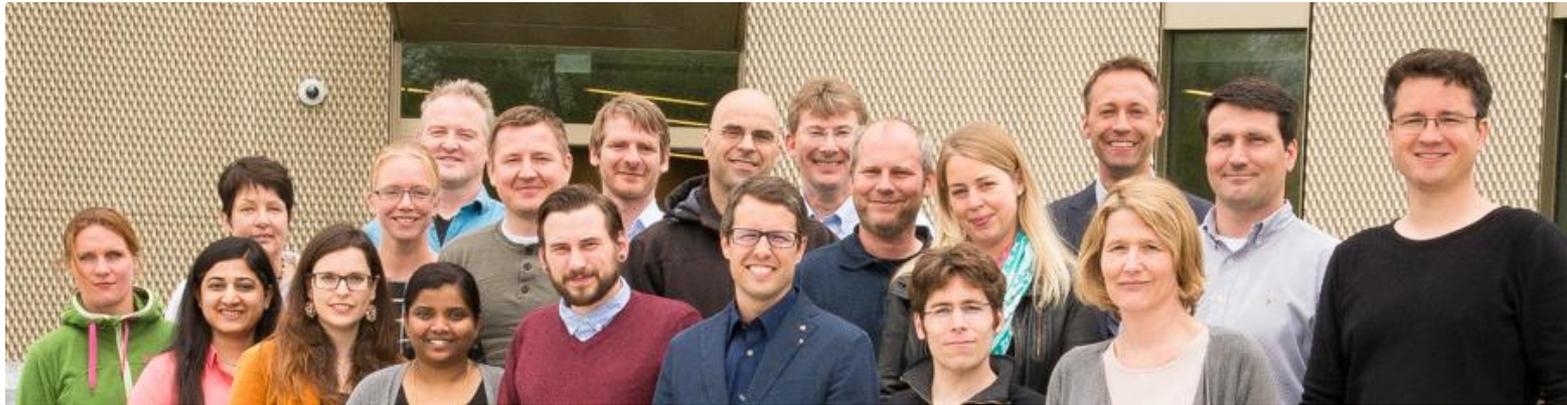


Christoph Gassner

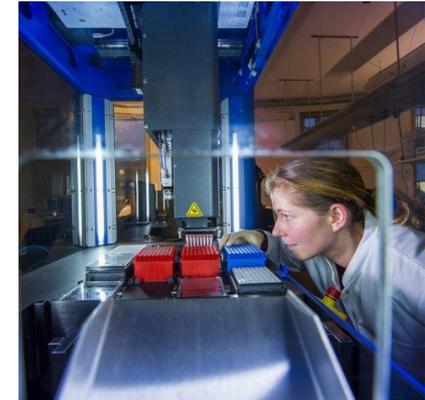


Michael Wittig

Acknowledgements



BLUTSPENDE ZÜRICH
| | | | |



a.franke@mucosa.de