Forschungsdatenmanagement biomedizinischer Genomdaten

SNP array und Next-Generation Sequenzierungs Daten

Michael Wittig

ikmb
Institut für Klinische Molekularbiologie

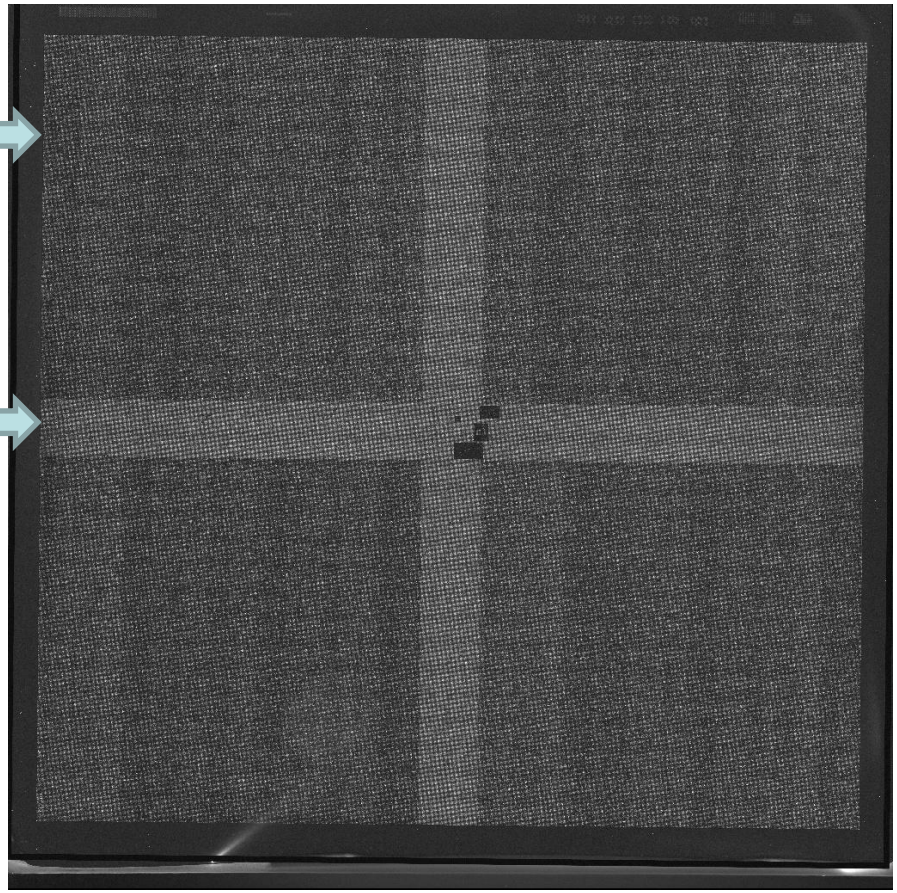
LABiMi
IF

SNP array Daten

Bsp: Affymetrix Genome Wide Human SNP array 6.0

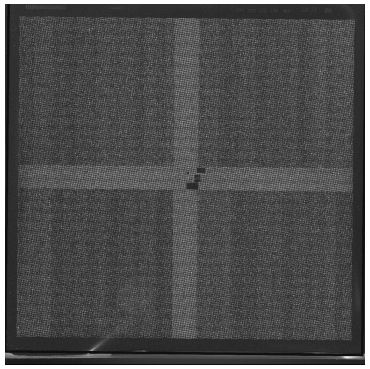
- 934,968 SNPs

- 746,000 CNV probes



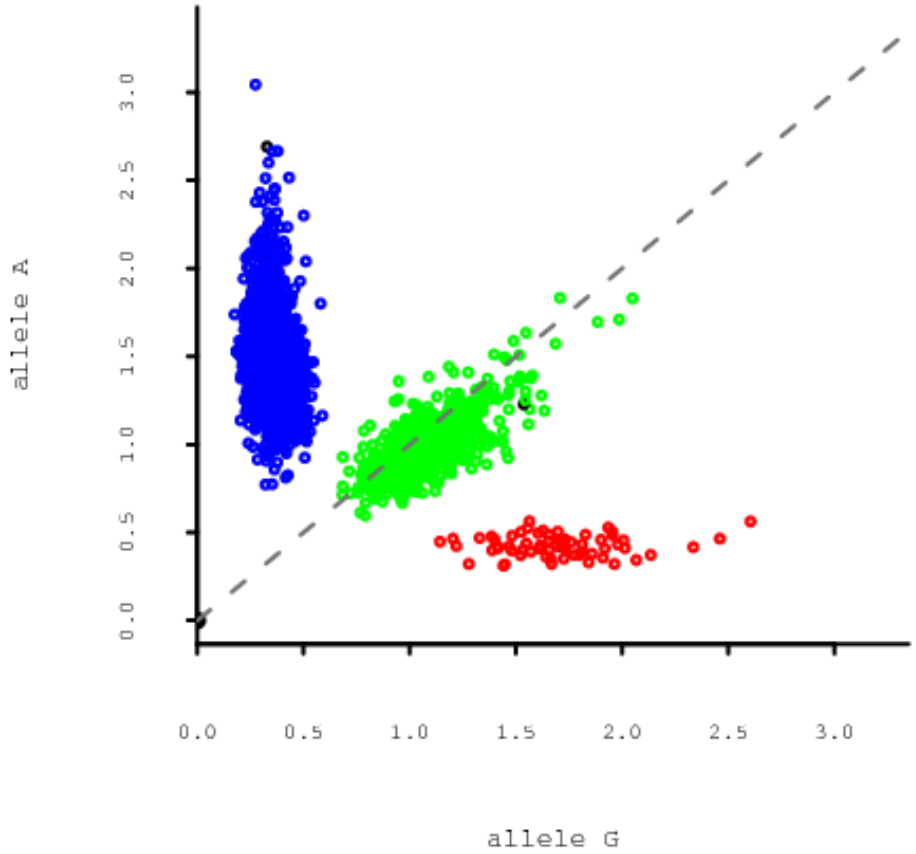
SNP array Daten

Bsp: Affymetrix Genome Wide Human SNP array 6.0



- 934,968 SNPs
- 946,000 SV probe sets

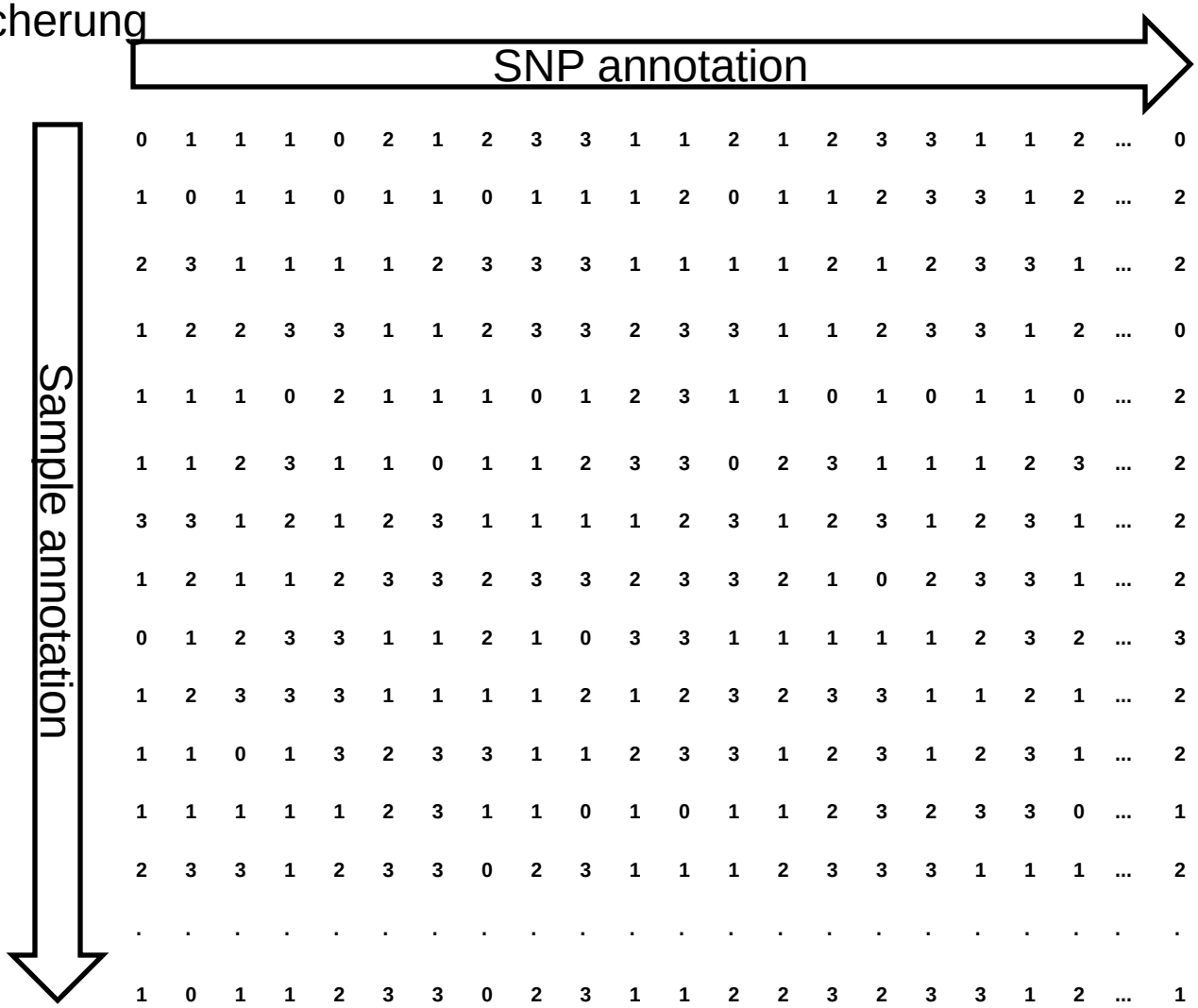
genotype AA
genotype AG
genotype GG



SNP array Daten

Gentotyp Speicherung

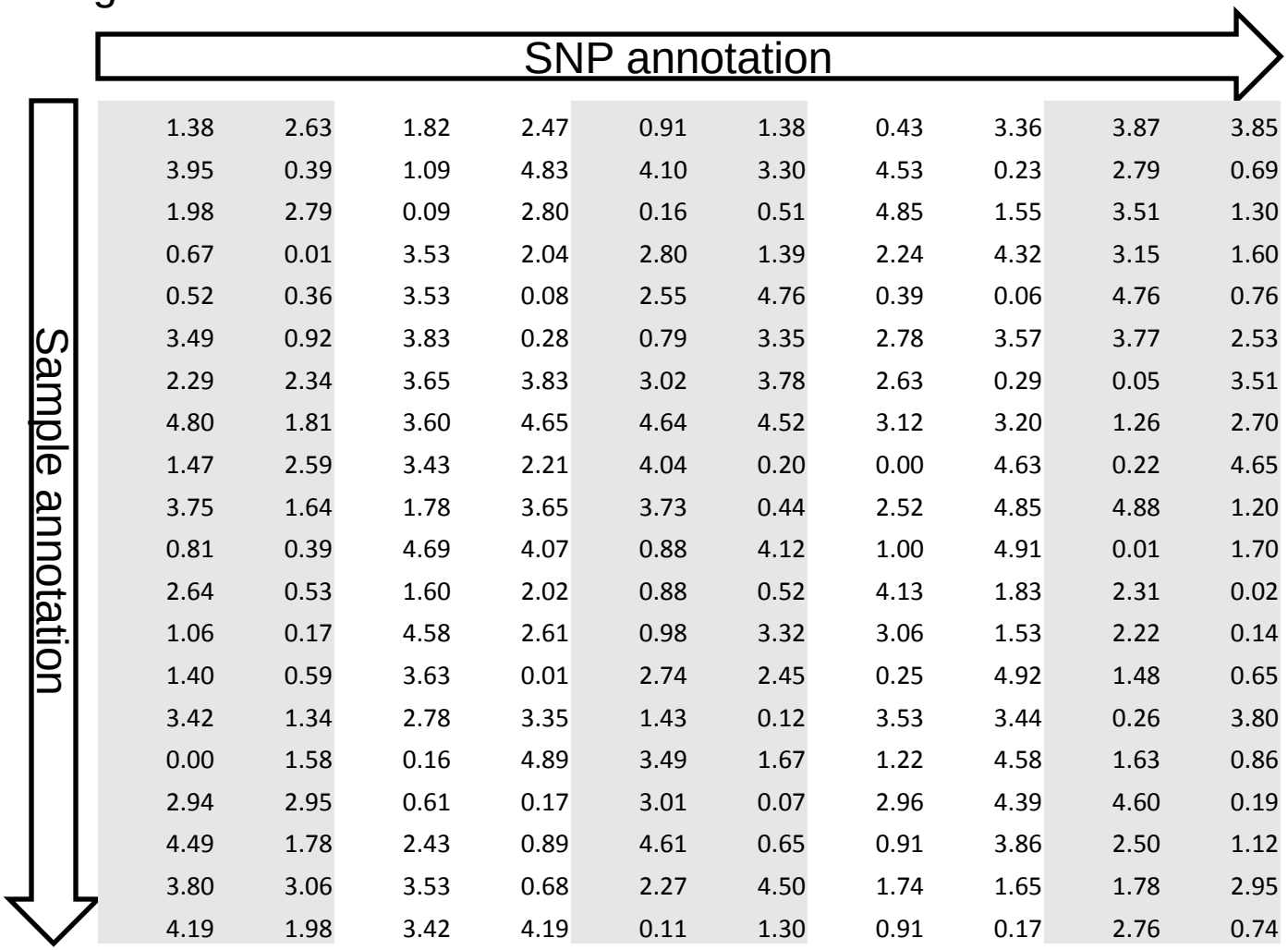
0 – NoCall
1 – Hom. A
2 – Heterozygot
3 – Hom. B



SNP array Daten

Intensitäten Speicherung

- 1 float/value
-> 2 float/SNP

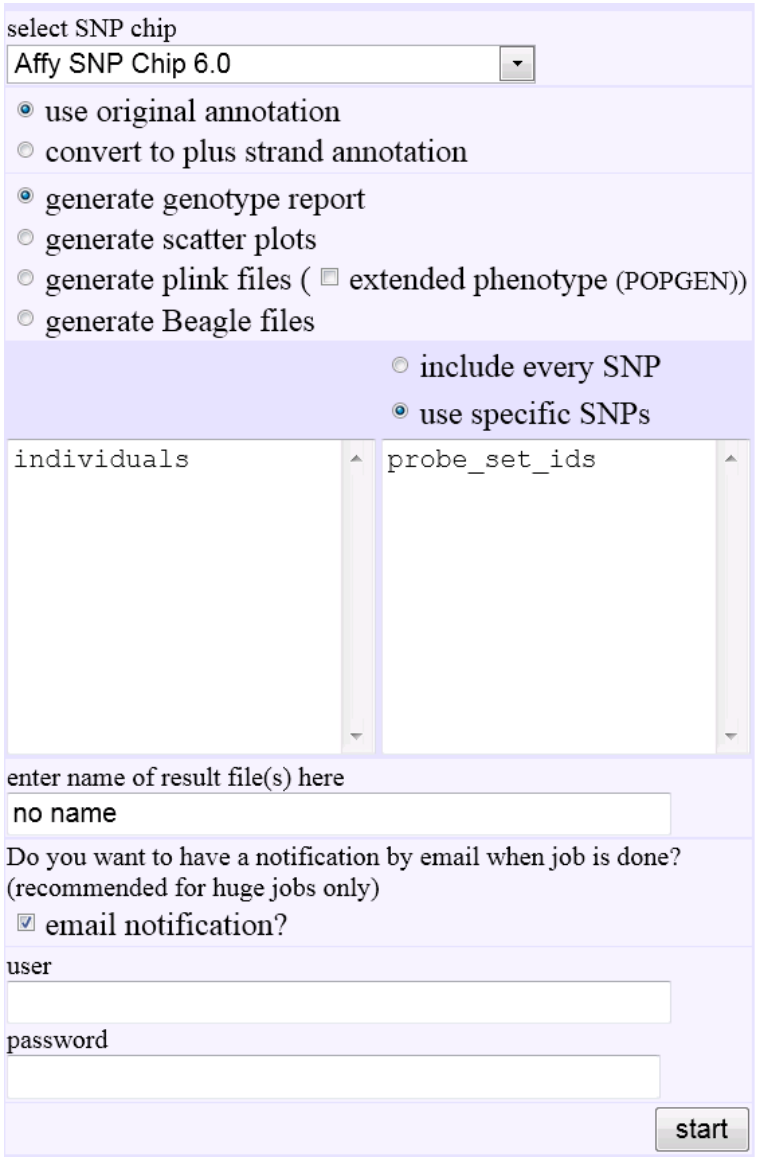


1.38	2.63	1.82	2.47	0.91	1.38	0.43	3.36	3.87	3.85
3.95	0.39	1.09	4.83	4.10	3.30	4.53	0.23	2.79	0.69
1.98	2.79	0.09	2.80	0.16	0.51	4.85	1.55	3.51	1.30
0.67	0.01	3.53	2.04	2.80	1.39	2.24	4.32	3.15	1.60
0.52	0.36	3.53	0.08	2.55	4.76	0.39	0.06	4.76	0.76
3.49	0.92	3.83	0.28	0.79	3.35	2.78	3.57	3.77	2.53
2.29	2.34	3.65	3.83	3.02	3.78	2.63	0.29	0.05	3.51
4.80	1.81	3.60	4.65	4.64	4.52	3.12	3.20	1.26	2.70
1.47	2.59	3.43	2.21	4.04	0.20	0.00	4.63	0.22	4.65
3.75	1.64	1.78	3.65	3.73	0.44	2.52	4.85	4.88	1.20
0.81	0.39	4.69	4.07	0.88	4.12	1.00	4.91	0.01	1.70
2.64	0.53	1.60	2.02	0.88	0.52	4.13	1.83	2.31	0.02
1.06	0.17	4.58	2.61	0.98	3.32	3.06	1.53	2.22	0.14
1.40	0.59	3.63	0.01	2.74	2.45	0.25	4.92	1.48	0.65
3.42	1.34	2.78	3.35	1.43	0.12	3.53	3.44	0.26	3.80
0.00	1.58	0.16	4.89	3.49	1.67	1.22	4.58	1.63	0.86
2.94	2.95	0.61	0.17	3.01	0.07	2.96	4.39	4.60	0.19
4.49	1.78	2.43	0.89	4.61	0.65	0.91	3.86	2.50	1.12
3.80	3.06	3.53	0.68	2.27	4.50	1.74	1.65	1.78	2.95
4.19	1.98	3.42	4.19	0.11	1.30	0.91	0.17	2.76	0.74

SNP array Daten

SNP Array Data Interface

- 8 MB/sample
- 553 GB am IKMB



SNP array Daten

SNP Array Data Interface

SNP annotation
Sarc_Immunochip.bed
Sarc_Immunochip.bim
Sarc_Immunochip.fam
Files are Individual major mode. Conversion to default (SNP major mode), use:
--bfile
--make-bed
--out
--allow-no-sex (optional)
If you get errors, try plink v1.02 for conversion ...
11:31:15 START PLINK-FILES a.fischer@ikmb.uni-kiel.de 2eda3d6e-bdd6-4f67-acb4-88d5e90344d5
USE ORIGINAL annotation, database iScan_Immuno_v2, build build36

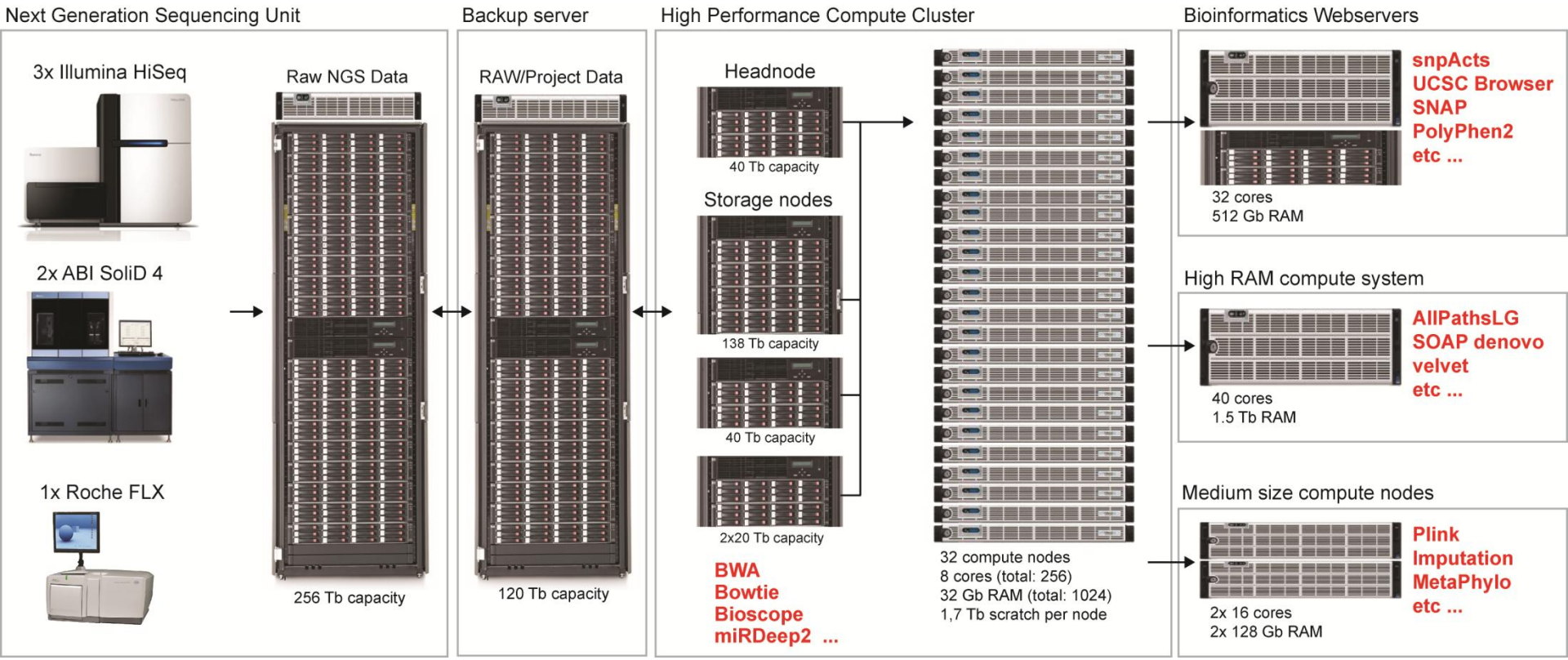
Unknown individual 'Controls' skipped.
Strand information: 102115 plus strand SNPs, 93617 minus strand SNPs and 792 SNPs without strand information.
results remain until Thursday, Mar 29 11:31 AM
11:31:29 END JOB

SNP array Daten

Zusammenfassung

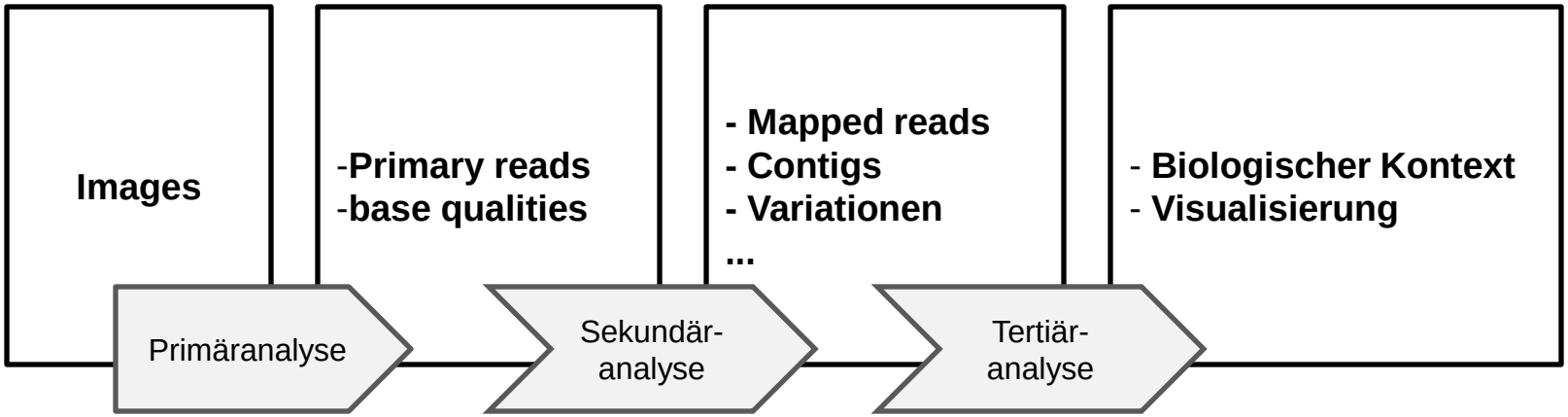
- Speicherung der Originaldaten (Affy6.0 z.Bsp. 30MB/array)
- Genotypbestimmung durch aktuellen calling Algorithmus
- QC nach Herstellerangaben (Affy6.0 z.B. call rate, contrastQC)
- Verfügbarkeit über hausinterne Datenbank im Plink Format
- SNP spezifische Qualitätskontrolle über scatter plots

NGS IT Infrastuktur



Bsp HiSeq: 2 flow cells/machine
1 flow cell run/10 days
160.000.000 Paired reads/lane (8 lanes/flow cell)
256.000.000.000 bases/flow cell in 10 days
1.536.000.000.000 bases/10 days

NGS Analysis Workflow



NGS quality control

fastQC

Basic Statistics

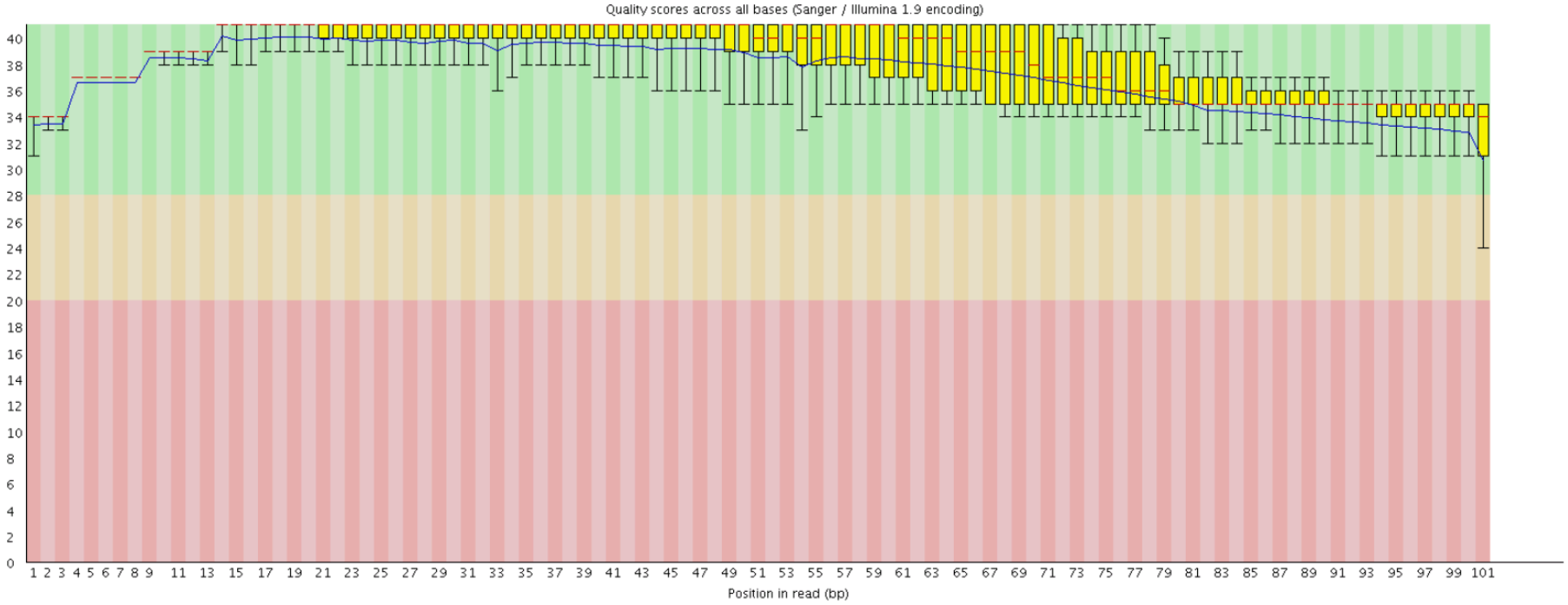
Measure	Value
Filename	A0850_ACTTGA_L006_R2_001.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	6525433
Sequence length	101
%GC	44

Back to summary

NGS quality control

fastQC

Per base sequence quality

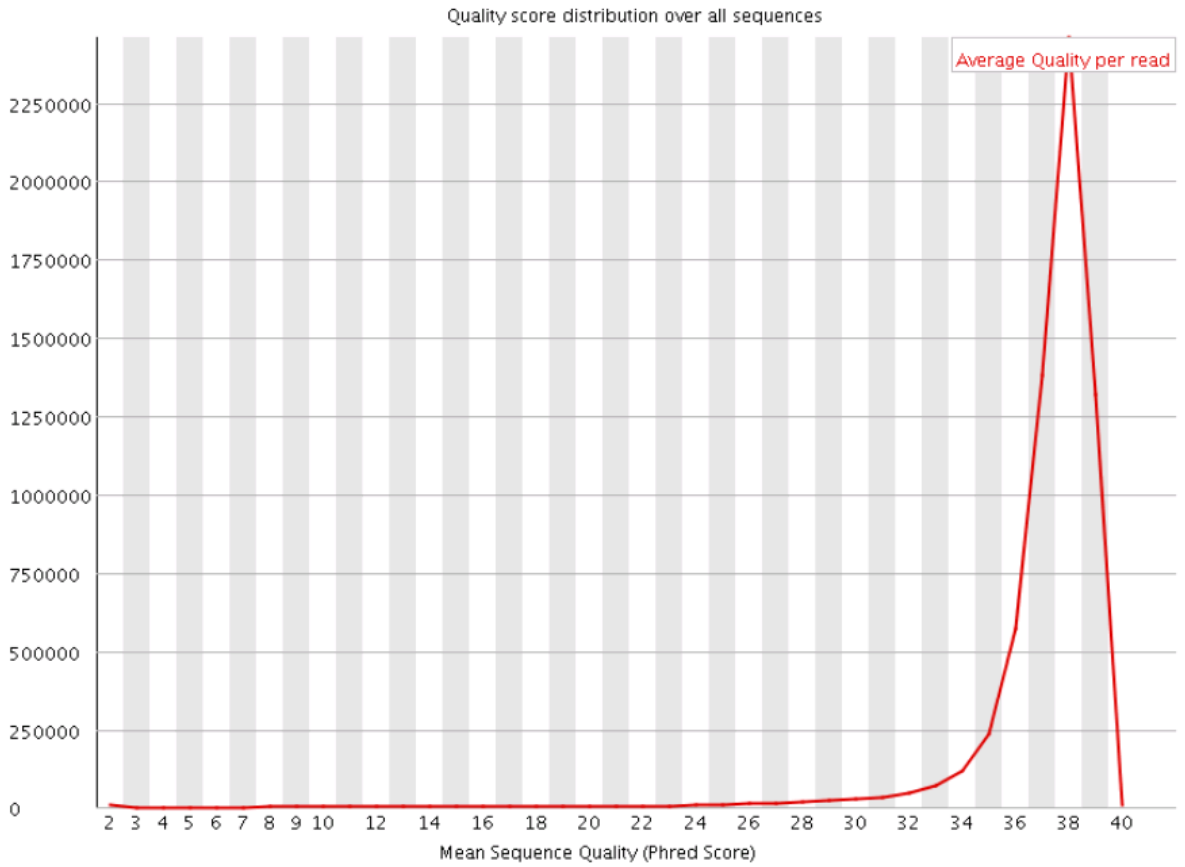


Back to summary

NGS quality control

fastQC

Per sequence quality scores

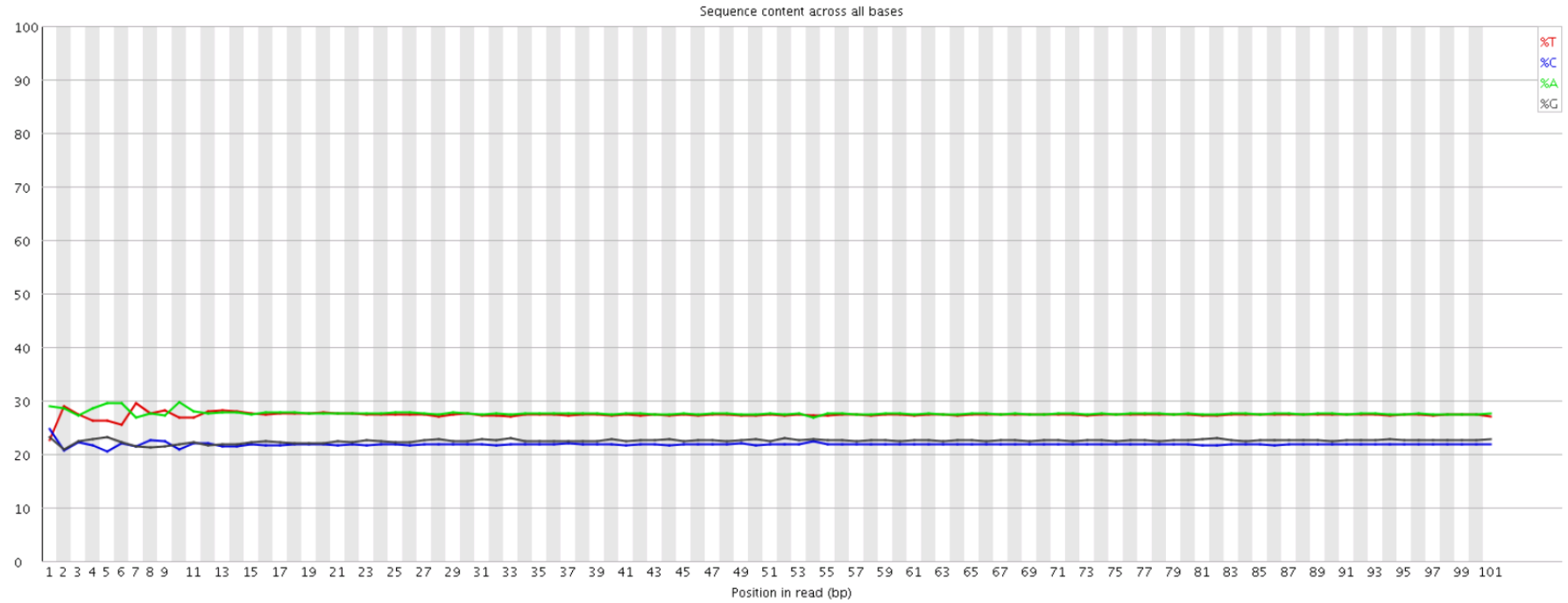


Back to summary

NGS quality control

fastQC

Per base sequence content



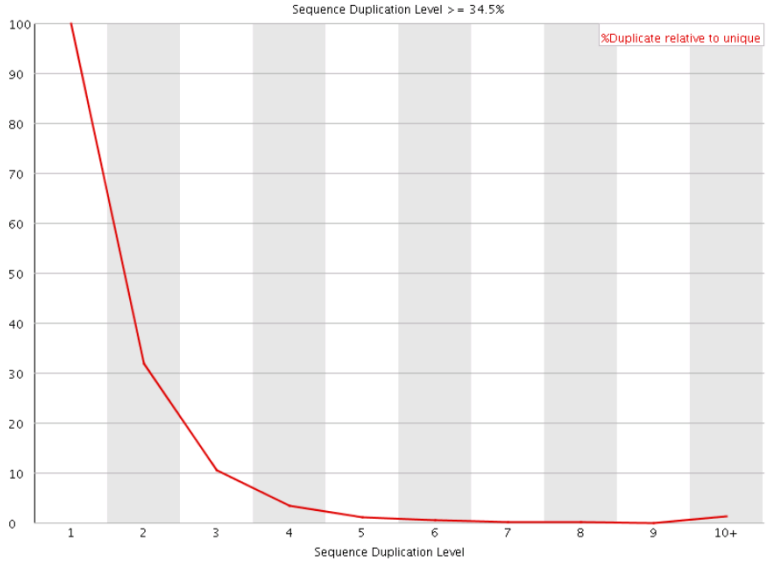
Back to summary

NGS quality control

fastQC

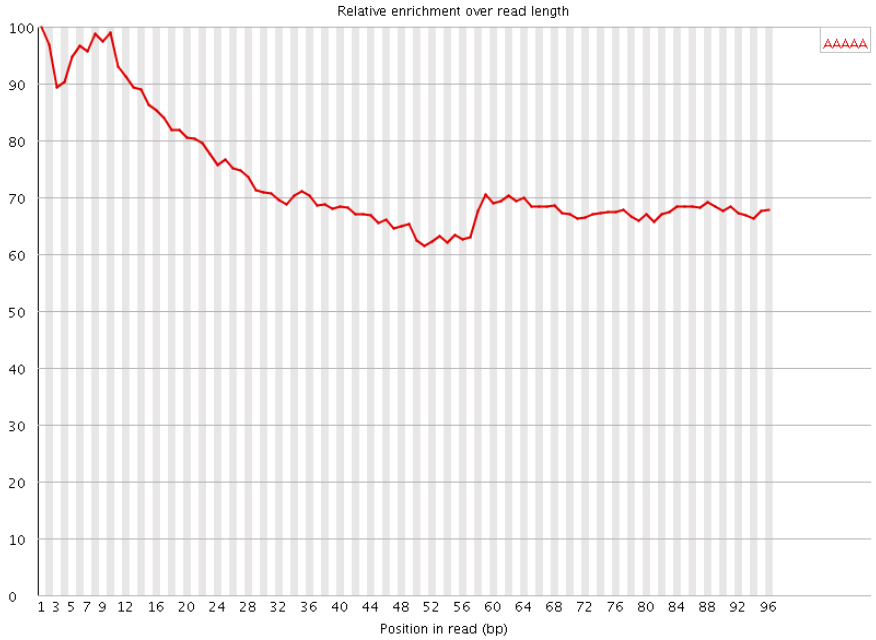
- Per Base GC content
- Per Sequence GC content
- Per Base N content
- Sequence length distribution

Sequence Duplication Levels



Back to summary

Kmer Content



Sequence	Count	Obs/Exp Overall	Obs/Exp Max	Max Obs/Exp Position
AAAAA	3175595	3.0460386	4.160364	1

Back to summary

NGS quality control

Metadaten / Projektbezogen

run_date
machine
flowcell
flowcell_id
run_id
lane
barcode
project
sample
project type
reference
description
#PF-reads
#bp(Mb)
mean quality PF
%Q30 bases
%PF
%raw cluster
%perfect index
%duplicates R1
%GC R1

avg. mean quality per base R1
avg. quality per base R1
#low qual positions R1
#pos with %N>=1 R1
max % N R1
avg. quality per sequence R1
%mapped R1
%duplicates R2
%GC R2
avg.mean quality per base R2
avg. quality per base R2
#low qual positions R2
#pos with %N>=1 R2
max % N R2
avg. quality per sequence R2
%mapped R2
avg.insert_size
median_insert_size

NGS quality control

Metadaten / Sequenzierungslauf Bezogen

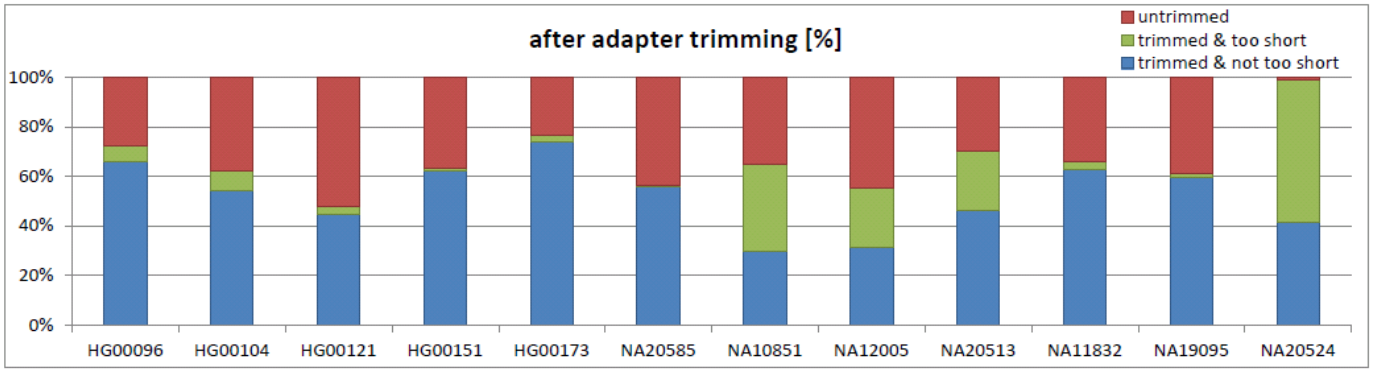
run_date
machine
flowcell
lane
run_id
flowcell_id
#total_reads
#PF_reads
%PF
%PF_reads_perfect_index
mean_per_base_qual(PF)
%Q30
%noIndex

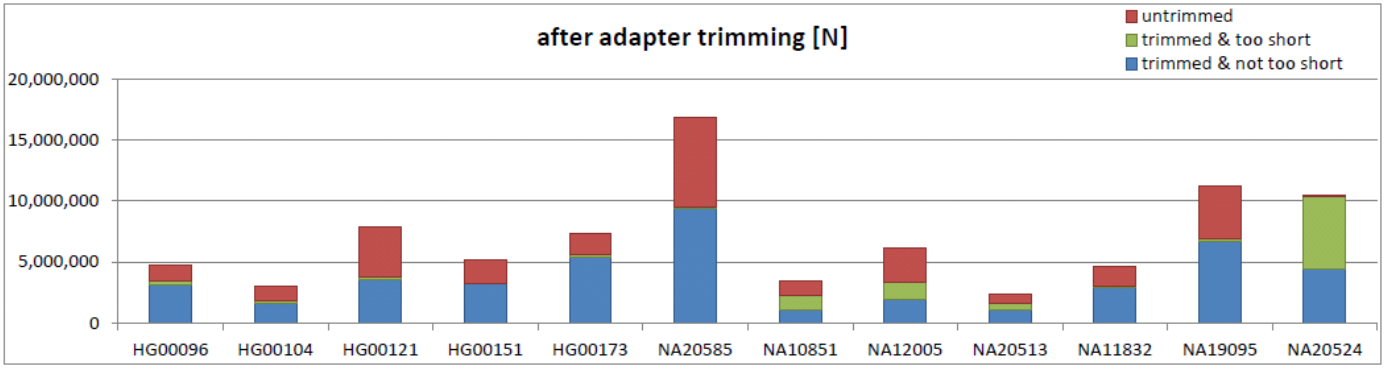
NGS quality control

Anwendungsspezifische QC

Adapter trimming results

	Lab 1			Lab 2			Lab 3			Lab 4		
	HG00096	HG00104	HG00121	HG00151	HG00173	NA20585	NA10851	NA12005	NA20513	NA11832	NA19095	NA20524
processed reads	4,701,889	3,040,116	7,903,163	5,178,053	7,314,599	16,878,747	3,489,043	6,113,322	2,384,113	4,599,349	11,285,616	10,507,737
trimmed reads	3,403,870	1,886,837	3,766,515	3,279,014	5,605,865	9,495,349	2,252,054	3,376,492	1,668,359	3,030,773	6,893,775	10,380,069
untrimmed	1,298,019	1,153,279	4,136,648	1,899,039	1,708,734	7,383,398	1,236,989	2,736,830	715,754	1,568,576	4,391,841	127,668
trimmed & too short	293,789	247,112	252,251	48,122	189,733	81,009	1,227,053	1,451,884	567,758	152,607	176,927	6,011,420
trimmed & not too short	3,110,081	1,639,725	3,514,264	3,230,892	5,416,132	9,414,340	1,025,001	1,924,608	1,100,601	2,878,166	6,716,848	4,368,649





NGS

Zusammenfassung

- Speicherung der „primary reads“ und „quality values“
- Spiegelung des Datenspeichers
- fastQC vor Sekundäranalyse
- Eigene Skripte zur Erzeugung von Metadaten
- Projektspezifisch angepaßte Qualitätskontrolle (Beispiel war miRNA)