Qualitätsmanagement für Hochdurchsatz-Genotypisierung

Work Package 4

Daten- und Qualitätsmanagement von Replikationsdaten („2. Stufe“) in SNP-Genotypisierungsstudien

Informationsveranstaltung, 21.06.2010, Berlin

Arne Pfeufer, HMGU München

SPONSORED BY THE

Federal Ministry of Education and Research

Scientific Problem - outline

- **Need for replication genotyping („2nd stage“) in GWAS studies**
  - increased **„n“** may be necessary to meet demanded significance thresholds
  - **no other GWAS samples** with genome-wide genotypes available
  - QTL-Hits from **imputed SNPs** may need non-silico **validation**
- **Quality issues in replication genotyping**
  - Replication genotyping **platforms** are **more diverse** and than GWAS platforms
    - ABI TaqMan
    - Sequenom MALDI
    - Kbioscences KASPAR
  - Replication genotyping **platforms** are **less standardized** than GWAS platforms
    - **No SOPs** exist for replication assay design
    - wide range of **error possibilities** in replication genotyping
      + assay design errors
      + genotyping errors

Scientific Problem - concrete application

- **QT-IGC : international GWAS consortium for QT interval**
  - QTGEN (n=13,685) identified 11 QTLs for QT interval (Nature Genetics, 2009)
  - QTSCD (n=15,842) identified 10 QTLs for QT interval (Nature Genetics, 2009)
  - + added (n=23,103) GWAS samples from 13 studies
  - Sum (n=52,630) identifies 18 QTLs for QT interval
  - + added (n=37,827) replication samples from 16 studies
  - Sum (n=90,457) identifies 27 QTLs for QT interval

- **QT-IGC - replication partners**
  - Genotyped between 5 and 35 SNPs
  - Using 3 different platforms:
    - ABI TaqMan
    - Sequenom MALDI
    - Kbioscences KASPAR

Scientific Problem - concrete application

- **QT-IGC : n=37,827 replication samples from 16 studies**

Replication cohorts	n	SNP platform	No of SNPs
British Regional Heart Survey/	3811	Kbioscience - KASPAR	5
Bruneck	629	Sequenom - MALDI	35
CARLA	1550	ABI - TaqMan	10
Cyprus	804	Sequenom - MALDI	35
FASTCARD	1305	ABI - TaqMan	30
Galicia	797	Sequenom - MALDI	35
Intergene-Adonix	2778	Kbioscience - KASPAR	10
MONICA-Prague	289	Sequenom - MALDI	35
PREVEND	7385	Kbioscience - KASPAR	5
SAPHIR	1288	Sequenom - MALDI	35
Whitehall-II	4510	Kbioscience - KASPAR	5
Young Finns	2008	ABI - TaqMan	20
Health2000	3128	Sequenom - MALDI	35
HNR	4469	Sequenom - MALDI	35
KORA-F3	908	Sequenom - MALDI	35
KORA-S4	2168	Sequenom - MALDI	35
Sum :	37827		

SNP_ID	CHR	CODED_ALLELE	NONCODED_ALL	CODED_ALLELE_FREQ	N_EFF	BETA_FIXED	SE_FIXED	Z_FIXED	P_FIXED	locus	new	Priority	BEST_5	BEST_15	BEST_20	BEST_30	BEST_36	#SNP rs#
rs121438	1	T	C	0.246	46210.1	3.399	0.139	24.427	8.84E-132	1	.	.	.	.	.	.	.	.
rs111537	6	T	C	0.51	46465.5	-1.602	0.119	-13.479	2.09E-41	2	.	.	.	.	.	.	.	.
rs37060	16	A	G	0.254	47983.7	-1.812	0.135	-13.456	2.84E-41	3	.	.	.	.	.	.	.	.
rs207423	11	T	C	0.19	43762.8	1.837	0.156	11.757	6.53E-32	4	.	.	.	.	.	.	.	.
rs296886	7	T	C	0.249	46629.5	-1.461	0.14	-10.448	1.50E-25	5	.	.	.	.	.	.	.	.
rs846111	1	C	G	0.287	27060.7	1.64	0.162	10.093	5.95E-24	6	.	.	.	.	.	.	.	.
rs109190	1	C	A	0.127	47043.6	-1.712	0.179	-9.585	9.22E-22	7	.	.	.	.	.	.	.	.
rs804960	16	T	C	0.501	36389.1	1.225	0.132	9.285	1.62E-20	8	.	.	.	.	.	.	.	.
rs139651	17	C	G	0.535	47690.2	-1.068	0.118	-9.054	1.38E-19	9	.	.	.	.	.	.	.	.
rs679324	3	A	G	0.322	45956.6	-1.076	0.129	-8.372	5.66E-17	10	.	.	.	.	.	.	.	.
rs313596	17	G	A	0.49	45327.3	-0.962	0.121	-7.96	1.72E-15	11	.	.	.	.	.	.	.	.
rs229863	1	T	C	0.495	42654	0.908	0.124	7.302	2.84E-13	12	1	5	.	.	.	30	36	rs229863
rs241405	15	A	T	0.461	47644.4	0.782	0.119	6.58	4.71E-11	13	2	5	.	.	.	30	36	rs241405
rs246185	16	C	T	0.329	36372.8	0.879	0.14	6.282	3.35E-10	14	3	5	.	.	.	30	36	rs246185
rs445277	8	G	C	0.408	44484.6	-0.761	0.122	-6.229	4.68E-10	15	4	5	.	.	.	30	36	rs445277
rs930350	17	G	C	0.44	46987.8	-0.728	0.12	-6.053	1.42E-09	16	5	5	.	.	.	30	36	rs930350
rs116850	2	G	C	0.369	45932.2	0.748	0.124	6.025	1.69E-09	17	6	5	.	.	.	30	36	rs116850
rs719548	16	C	G	0.215	33835.9	0.893	0.162	5.499	3.82E-08	18	7	4	.	.	.	30	36	rs719548
rs756114	2	C	T	0.416	47068	-0.656	0.12	-5.456	4.87E-08	19	8	4	.	.	.	30	36	rs756114
rs938291	2	G	C	0.386	45699.3	0.666	0.123	5.399	6.72E-08	20	9	1	5	15	20	30	36	rs938291
rs9920	7	C	T	0.094	44716.4	1.098	0.207	5.299	1.16E-07	21	10	1	5	15	20	30	36	rs9920
rs174583	11	T	C	0.333	45550.8	-0.667	0.127	-5.268	1.38E-07	22	11	1	5	15	20	30	36	rs174583
rs174536	.	.	.	.	.	.	.	.	.	.	11	.	.	.	.	.	.	rs174536
rs385706	4	A	T	0.45	47403.9	-0.619	0.119	-5.205	1.94E-07	23	12	1	5	15	20	30	36	rs385706
rs295140	2	T	C	0.416	47287	0.62	0.12	5.162	2.45E-07	24	13	1	5	15	20	30	36	rs295140
rs776582	6	G	C	0.385	44854.2	0.638	0.124	5.136	2.80E-07	25	14	2	.	15	20	30	36	rs776582
rs302644	12	C	T	0.355	47483.2	0.631	0.124	5.083	3.72E-07	26	15	2	.	15	20	30	36	rs302644
rs404321	7	A	G	0.003	5900	13.679	2.772	4.935	8.00E-07	27	16	2	.	15	20	30	36	rs404321
rs227390	14	T	C	0.348	29919.9	0.757	0.156	4.837	1.32E-06	28	17	2	.	15	20	30	36	rs227390
rs449391	8	T	C	0.212	46472.8	0.701	0.145	4.821	1.43E-06	29	18	2	.	15	20	30	36	rs449391
rs169715	15	C	T	0.049	18629.5	-1.921	0.399	-4.81	1.51E-06	30	19	2	.	15	20	30	36	rs169715
rs169284	9	T	C	0.013	23764.9	3.528	0.734	4.806	1.54E-06	31	20	2	.	15	20	30	36	rs169284
rs728478	17	G	A	0.447	47137.1	-0.562	0.119	-4.728	2.26E-06	32	21	2	.	15	20	30	36	rs728478
rs168706	5	T	C	0.399	35443.9	0.637	0.136	4.672	2.98E-06	33	22	2	.	15	20	30	36	rs168706
rs196110	8	T	C	0.337	41486.7	0.611	0.132	4.64	3.48E-06	34	23	2	.	15	20	30	36	rs196110
rs259307	4	A	G	0.388	44820.1	-0.575	0.124	-4.63	3.65E-06	35	24	3	.	.	20	30	36	rs259307
rs101252	9	A	G	0.023	7277.1	4.409	0.953	4.628	3.70E-06	36	25	3	.	.	20	30	36	rs101252
rs177515	1	A	T	0.429	47257.7	0.552	0.12	4.617	3.90E-06	37	26	3	.	.	20	30	36	rs177515
rs619540	2	C	T	0.145	26598.6	0.965	0.209	4.612	3.98E-06	38	27	3	.	.	20	30	36	rs619540
rs124425	15	A	C	0.078	43721.8	1.066	0.232	4.595	4.33E-06	39	28	3	.	.	20	30	36	rs124425
rs780897	7	G	A	0.394	38137.6	0.603	0.132	4.57	4.88E-06	40	29	4	.	.	.	30	36	rs780897
rs780653	.	.	.	.	.	.	.	.	.	.	29	.	.	.	.	.	.	rs780653
rs172172	13	T	C	0.123	46285.8	-0.839	0.185	-4.543	5.56E-06	41	30	4	.	.	.	30	36	rs172172
rs134031	2	G	A	0.31	34574.9	-0.656	0.145	-4.514	6.36E-06	42	31	4	.	.	.	.	36	rs134031
rs100409	5	A	G	0.131	47719.1	-0.786	0.175	-4.5	6.79E-06	43	32	4	.	.	.	.	36	rs100409
rs171327	7	G	T	0.009	24129.4	3.655	0.815	4.486	7.26E-06	44	33	4	.	.	.	.	36	rs171327
rs118681	17	A	G	0.062	41441.7	-1.208	0.27	-4.481	7.44E-06	45	34	4	.	.	.	.	36	rs118681
rs432995	.	.	.	.	.	.	.	.	.	.	34	.	.	.	.	.	.	rs432995
rs100621	5	T	C	0.007	7857.6	7.165	1.604	4.466	7.96E-06	46	35	4	.	.	.	.	36	rs100621
rs284163	6	T	C	0.04	39615.7	1.387	0.311	4.46	8.19E-06	47	36	4	.	.	.	.	36	rs284163
rs742691	3	G	A	0.432	30733.1	-0.657	0.147	-4.454	8.44E-06	48	37	5	.	.	.	.	36	rs742691
rs112334	11	G	A	0.083	40104.9	1.026	0.232	4.432	9.33E-06	49	38	5	.	.	.	.	36	rs112334
rs227412	14	C	G	0.379	46675.6	0.542	0.122	4.43	9.42E-06	50	39	5	.	.	.	.	36	rs227412
rs690900	6	A	G	0.115	42681.9	0.855	0.193	4.421	9.83E-06	51	40	5	.	.	.	.	36	rs690900
rs654979	17	C	G	0.17	43546.9	0.72	0.163	4.418	9.94E-06	52	41	5	.	.	.	.	36	rs654979

Work Package Description

- **Projected goals**
  - Quality control of **SNP replication genotyping** (GWAS „2nd stage“)
    - QC for
      - **assay design**
      - **strand-sign issue**
      - **HWE**

    - NOT: QC of allele calling clustering in 2nd stage genotyping (needs too much raw data)
    - NOT: QC of imputed genotype data
    - NOT: QC of beta effect estimatior sign (done at the metaanalysis stage)

Work Package Description

- **Targeted Results – specific**
  - Ensuring genotyping of the **correct SNP**
  - Ensuring the correct **strand annotation** (+/-)
    - especially important for **homonymous SNPs** (A/T, C/G)
  - Ensuring **classical QC** metrices: callrate, p(HWE), MAF
    - Comparison of data to HapMap data

- **Deliverables**
  - Software tool **„RepliCheckSNP“** that
    - QC's replication **genotyping assays for their accurate performance**
      - **–** no tools exist currently for this task

    NOT: Checks replication **genotypes** – solvalbe by existing tools (plink)
    NOT: Checks replication **association data** – task of metaanalysis project statistician

RepliCheckSNP Software tool

- Required **input information** – for each individual SNP
  - **SNP** targeted for genotyping
  - **Genome assembly** used
  - **Position** of that SNP in genome assembly
  - **Used strand orientation** of that SNP relative in genome assembly
  - **Sequence flanks** used for assay design
  - Oligonucleotides used to **amplify** genomic segment
  - Oligonucleotide(s) used to **probe** SNP (3 existing assay methods)
    a. probe covering SNP (e.g. TaqMan)
    b. probe ending on SNP (e.g. ligation assays)
    c. probe ending before SNP (e.g. Sequenom, Minisequencing)
  - **Allelic variants** targeted by assay (A,C,G,T or subset only)
  - Genotyping **results statistic** (AA, Aa, aa)

RepliCheckSNP Software tool

- **Approach - „RepliCheckSNP“ Software tool**
  - **Analysis performed** by the tool – for each individual SNP
    - Report **presence/absence** of SNP in latest genome assemblies
    - Automatically **align sequence flanks** to genome assembly
    - Ensure that **sequence flanks** reported map to genome assembly
    - **Compare reported position** with mapped position
    - Check if **oligonucleotides** used to **amplify** genomic segment bind uniquely to genome
    - Based on the **oligonucleotide** used to **probe** the SNP (3 existing assay methods) determine:
      - correct assaying of desired base at desired position
      - Strand orientation of the assay results
    - determine CR, p(HWE) and MAF and compare to dSNP, HapMap etc.
- **Comparison to existing results report sheets**
  - Results report sheet of the CHARGE consortium
- **Realization of the Software solution „RepliCheckSNP“**
  - Implementation in JAVA code to enhance trans-platform portability

Results Data Structure - CHARGE consortium

Variable name	Description
SNPID	SNP ID as rs number
chr	chromosome number. Use symbols X, XY, Y and mt for non-autosomal markers.
position	physical position for the reference sequence (build 35 strongly preferred)
coded_all	coded allele, also called modeled allele (in example of A/G SNP in which AA=0, AG=1 and GG=2, the coded allele is G)
noncoded_all	the alternate allele
strand_genome	+ or -, representing either the positive/forward strand or the negative/reverse strand of the human genome reference sequence; to clarify which strand the coded_all and noncoded_all are on
beta	beta estimate from genotype-phenotype association, at least 5 decimal places -- NA if not available
SE	standard error of beta estimate, to at least 5 decimal places -- NA if not available
Pval	p-value of test statistic, here just as a double check -- NA if not available
AF_coded_all	allele frequency for the coded allele -- NA if not available
HWE_pval	exact test Hardy-Weinberg equilibrium p-value -- only directly typed SNPs, NA for imputed
callrate	genotyping callrate after exclusions
n_total	total sample with phenotype and genotype for SNP
imputed	Only in case imputed data are used, otherwise NA --1/0 coding; 1=imputed SNP, 0=if directly typed
used_for_imp	Only in case imputed data are used, otherwise NA --1/0 coding; 1=used for imputation, 0=not used for imputation
oevar_imp	Only in case imputed data are used, otherwise NA --observed divided by expected variance for imputed allele dosage
avpostprob	Only in case imputed data are used, otherwise NA --average posterior probability for imputed SNP allele dosage* (applies to best- guess genotype imputation)

Flowchart for „RepliCheckSNP“ Software tool

A. Input:

1. One input file per study
2. File contains assay design and genotyping results information from each study

B. Web based query:

1. SNP flanking sequences from UCSC
2. SNP genotyping information from HapMart on chosen population(using wget)

C. Output - Three software routines generate three output files per study:

1. checkPOS:
   - Presence of SNP in chosen genome assembly (currently NCBI Build 37)
   - Correctness of given position to actual position of sequence in the genome
2. checkBLAT: Alignment of SNP probe sequence to chosen genome assembly
3. checkHWE:
   - Compares MAF and p(HWE) to HapMap using chosen population (e.g.CEU)
   - Detects strand (+/-) switches (e.g. C/T -> G/A) and coded allele switches (e.g. C/T -> T/C)

Flowchart for „RepliCheckSNP“ Software tool

A. Input:

1. One input file per study
2. File contains assay design and genotyping results information from each study

B. Web based query:

1. SNP flanking sequences from UCSC
2. SNP genotyping information from HapMart on chosen population(using wget)

C. Output - Three software routines generate three output files per study:

1. checkPOS:
   - Presence of SNP in chosen genome assembly (currently NCBI Build 37)
   - Correctness of given position to actual position of sequence in the genome
2. checkBLAT: Alignment of SNP probe sequence to chosen genome assembly
3. checkHWE:
   - Compares MAF and p(HWE) to HapMap using chosen population (e.g.CEU)
   - Detects strand (+/-) switches (e.g. C/T -> G/A) and coded allele switches (e.g. C/T -> T/C)

Input file „RepliCheckSNP“ Software tool - Part 1

ALL Sentinel_SNP_ID_as_rs_number	ALL SNP_replaced_by_proxy	ALL most_plausible_gene_in_region	ALL chromosome_number._Use_symbols_X,_XY,_Y_and_mt_for_non-autosoma	ALL physical_position_for_the_reference_sequence_in_MB	. .	ALL study_in_which_SNP_was_typed	TAQMAN_ONLY Genotyped_SNP_ID_as_rs_number	ALL NCBI_genome_build_used,_i.e._to_which_"chr"_and_"pos"_above_do_refer	TAQMAN_ONLY Sequence_of_probe_/_sequencing_primer_/_etc._used_to_assay_the_SNP
Sentinel_SNP_ID	SNP_proxy_necessary	LOCUS	CHR	POSITION_MB	Band	STUDY	Genotyped_SNP_ID_TAQMAN	NCBI_build_used	PROBE_sequence_TAQMAN
rs174583	N	FADS2	11	61.4	11q12.2	CARLA	rs174583	build36	AGCTTGCCTGGCCCTGAGCCTGAAG[C/T]GGCCTGAGAACCTGGTCTCTGTCCA
rs2273905	N	ANKRD9	14	102.0	14q32.31	CARLA	rs2273905	build36	TTACACGTTAGCTCCTGGGAGGAGA[C/T]AGGAGGGTGAAAACAACCTGGAGAC
rs295140	N	LOC26010	2	200.9	2q33.1	CARLA	rs295140	build36	TGCCATTTTCATTATATCCTCCTTC[C/T]TTCCCTGAACCAGTGTCCTGTCTTA
rs3026445	N	ATP2A2	12	109.2	12q24.11	CARLA	rs3026445	build36	ATTCTCTTCTTGATTACCAACTTTG[C/T]TCTAAATGTAACACATCGTATGGTT
rs3857067	N	SMARCAD1	4	95.2	4q22.2	CARLA	rs3857067	build36	TCTATTTTAATTATATGGAAAGGTA[A/T]TTGCATCATCTCAATTAAGTTGTAT
rs4493911	N	LOXL2	8	23.2	8p21.3	CARLA	rs4493911	build36	CCGACAGGGAAGTGGCTGTTCCTCT[C/T]TGCTGGGATGCTTTCCCCCGGGGAG
rs7765828	N	GMPR	6	16.4	6p22.3	CARLA	rs7765828	build36	AAGACACTTGATTTCTGATCTTAGA[C/G]CAGACAGGCGGGAGGTGAAGCTCTG
rs938291	N	SP3	2	174.5	2q31.1	CARLA	rs938291	build36	TTATCAACTGAAACTGAAATACCTA[C/G]ACTTTAACAAGATGTTAAGTATGTA
rs9920	N	CAV1	7	116.0	7q31.2	CARLA	rs9920	build36	CTCCCTGAAGACCAAAATTAGAATA[C/T]CCATGACCTAGTTTTCCATGCGTGT

Input file „RepliCheckSNP“ Software tool – Part 2

ALL Sentinel_SNP_ID _as_rs_number	TAQMAN_ONLY Sequence_of_probe_/_sequencing_primer_/_etc._used_to_assay_the_SNP	TAQMAN_ONLY F_or_R,_representing_either_the_positive/forward_strand_or_the_negative/reverse_strand_of_the_human_genome_from_which_th	ALL .	ALL .	ALL number_of_homozygotes_for_the_coded_allele	ALL number_of_heterozygotes	ALL number_of_homozygotes_for_the_noncoded_allele	ALL beta_estimate_from_genotype-phenotype_association,_at_least_5_decimal_places_--	ALL standard_error_of_beta_estimate,_to_at_least_5_decimal_places_--_“NA”_if_not_available_	ALL p-value_of_test_statistic,_here_just_as_a_double_check_--_“NA”_if_not_available_
Sentinel_SNP_ID	**PROBE_sequence_TAQMAN**	**strand_genome_TAQMAN**	**coded_allele_genotyped_SNP**	**noncoded_or_reference_allele_genotyped_SNP**	**n_homozygous_coded_allele**	**n_heterozygous**	**n_homozygous_noncoded_allele**	**beta_Coded_Allele_additive_model**	**StdErr_Coded_Allele**	**P-value**
rs174583	AGCTTGCCTGGCCCTGAGCCTGAAG[C/T]GGC(	+	T	C	188	684	669	0.3203	0.71945193	0.6562
rs2273905	TTACACGTTAGCTCCTGGGAGGAGA[C/T]AGGA	+	T	C	185	744	615	0.2391	0.73141634	0.7438
rs295140	TGCCATTTTCATTATATCCTCCTTC[C/T]TTCCC	+	T	C	271	747	527	1.9	0.69216758	0.006127
rs3026445	ATTCTCTTCTTGATTACCAACTTTG[C/T]TCTAA	+	C	T	184	713	610	0.7276	0.74025842	0.3258
rs3857067	TCTATTTTAATTATATGGAAAGGTA[A/T]TTGCA	-	T	A	360	733	452	0.2242	0.67347552	0.7393
rs4493911	CCGACAGGGAAGTGGCTGTTCCTCT[C/T]TGCT	+	T	C	52	454	1032	-0.7775	0.89009731	0.3825
rs7765828	AAGACACTTGATTTCTGATCTTAGA[C/G]CAGA(	+	G	C	292	764	488	0.1077	0.69305019	0.8766
rs938291	TTATCAACTGAAACTGAAATACCTA[C/G]ACTTT	+	G	C	247	734	562	0.0124	0.69937958	0.9859
rs9920	CTCCCTGAAGACCAAAATTAGAATA[C/T]CCATG	+	C	T	10	289	1247	0.3938	1.16819935	0.7361

Flowchart of„RepliCheckSNP“ Software tool

A. Input:

1. One input file per study
2. File contains assay design and genotyping results information from each study

B. Web based query:

1. SNP flanking sequences - from UCSC
2. SNP genotyping information on chosen population(using wget) - from HapMart

C. Output - Three software routines generate three output files per study:

1. checkPOS:
   - Presence of SNP in chosen genome assembly (currently NCBI Build 37)
   - Correctness of given position to actual position of sequence in the genome
2. checkBLAT: Alignment of SNP probe sequence to chosen genome assembly
3. checkHWE:
   - Compares MAF and p(HWE) to HapMap using chosen population (e.g.CEU)
   - Detects strand (+/-) switches (e.g. C/T -> G/A) and coded allele switches (e.g. C/T -> T/C)

Input for „RepliCheckSNP“ - UCSC

SNP-flank data are retrieved from UCSC for each SNP by rs-Number - using BLAT:

program checkBlat Alignment

description: check of probeSequ Blat align

input: QT-IGC__QC_TaqMan_CARLA.xls

output: QT-IGC__QC_TaqMan_CARLA_checkBlat.xls

command logging:
url: http://genome.cse.ucsc.edu/cgi-bin/hgBlat?org=Human&db=hg18&type=DNA&sort=query,score&output=psl&hgsid=151659628&userSeq=AGCTTGCCTGGCCCTGAGCCTGAAGCGGCCTGAGAACCTGGTCTCTGTCCA

url=http://genome.cse.ucsc.edu/cgi-bin/hgc?hgsid=151678254&db=hg18&g=htcGetDna2&getDnaPos=chr11:6136630061366351&hgSeq.casing=upper&hgSeq.repMasking=lower&submit=get+DNA

url: http://genome.cse.ucsc.edu/cgi-bin/hgBlat?org=Human&db=hg18&type=DNA&sort=query,score&output=psl&hgsid=151659628&userSeq=TTACACGTTAGCTCCTGGGAGGAGACAGGAGGGTGAAAACAACCTGGAGAC

url=http://genome.cse.ucsc.edu/cgi-bin/hgc?hgsid=151678254&db=hg18&g=htcGetDna2&getDnaPos=chr14:102044726-102044777&hgSeq.casing=upper&hgSeq.repMasking=lower&submit=get+DNA

url: http://genome.cse.ucsc.edu/cgi-bin/hgBlat?org=Human&db=hg18&type=DNA&sort=query,score&output=psl&hgsid=151659628&userSeq=TGCCATTTTCATTATATCCTCCTTCCTTCCCTGAACCAGTGTCCTGTCTTA

url=http://genome.cse.ucsc.edu/cgi-bin/hgc?hgsid=151678254&db=hg18&g=htcGetDna2&getDnaPos=chr2:200868918-200868969&hgSeq.casing=upper&hgSeq.repMasking=lower&submit=get+DNA

...

</D

Input for „RepliCheckSNP“ - HapMart

SNP-assay data are retrieved from HapMart - using wget :

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE Query>
<Query  virtualSchemaName = "rel27_NCBI_Build36" formatter = "TSV" header = "0" uniqueRows = "0" count = "" datasetConfigVersion = "0.6" >
            <Dataset name = "hm27_variation_ceu" interface = "default" >
                        <Filter name = "marker_name" value = "rs174583,rs2273905,rs295140,rs3026445"/>
                        <Attribute name = "chrom" />
                        <Attribute name = "start" />
                        <Attribute name = "strand" />
                        <Attribute name = "marker1" />
                        <Attribute name = "ref_allele" />
                        <Attribute name = "ceu_id" />
                        <Attribute name = "other_allele" />
                        <Attribute name = "refhom_gtcount" />
                        <Attribute name = "het_gtcount" />
                        <Attribute name = "otherhom_gtcount" />
                        <Attribute name = "assay_id" />
            </D
```

HapMart returns the output:

chromosome	position	strand	marker id	reference allele	other allele	reference homozygote genotype count	heterozygote genotype count	other homozygote genotype count	genotyping platform
chr2	200868944	+	rs295140	T	C	19	58	36	Illumina_1M
chr2	174450854	+	rs938291	G	C	16	57	39	AFFY_6.0
chr4	95245457	+	rs3857067	T	A	31	60	22	AFFY_6.0
chr6	16402701	+	rs7765828	C	G	43	52	18	AFFY_6.0
chr7	115987328	+	rs9920	T	C	95	17	0	Illumina_1M
chr8	23229408	+	rs4493911	T	C	1	20	30	Perlegen
chr11	61366326	+	rs174583	C	T	49	49	15	AFFY_6.0
chr12	109207586	+	rs3026445	T	C	52	49	12	Illumina_1M
chr14	102044752	+	rs2273905	T	C	15	38	60	Illumina_1M

Flowchart of „RepliCheckSNP“ Software tool

A. Input:

1. One input file per study
2. File contains assay design and genotyping results information from each study

B. Web based query:

1. SNP flanking sequences from UCSC
2. SNP genotyping information from HapMart on chosen population(using wget)

C. Output - Three software routines generate three output files per study:

1. checkPOS:
   - Presence of SNP in chosen genome assembly (currently NCBI Build 37)
   - Correctness of given position to actual position of sequence in the genome
2. checkBLAT: Alignment of SNP probe sequence to chosen genome assembly
3. checkHWE:
   - Compares MAF and p(HWE) to HapMap using chosen population (e.g.CEU)
   - Detects strand (+/-) switches (e.g. C/T -> G/A) and coded allele switches (e.g. C/T -> T/C)

Output of „RepliCheckSNP“ - 1. checkPOS

Source	Validation	Sentinel_SNP_ID_as_rs_number	SNP_replaced_by_proxy	most_plausible_gene_in_region	chromosome_number._Use_symbols_X,_XY,_Y_and_mt_for_non-autosomal_markers._	physical_position_for_the_reference_sequence	.	study_in_which_SNP_was_typed	strand_genotyped_SNP	NCBI_genome_build_used,_i.e._to_which_"chr"_and_"pos"_above_do_refer
reported_data	Result	Sentinel_SNP_ID	SNP_proxy_necessary	LOCUS	CHR	Position SNP	Band	STUDY	strand_genome	BUILD_used_NCBI_or_USCS
input data		rs174583	N	FADS2	11	61366326	11q12.2	CARLA	+	build36
validation	ok				11	61366326			+	hg18
input data		rs2273905	N	ANKRD9	14	102044752	14q32.31	CARLA	+	build36
validation	ok				14	102044752			+	hg18
input data		rs295140	N	LOC26010	2	200868944	2q33.1	CARLA	+	build36
validation	ok				2	200868944			+	hg18
input data		rs3026445	N	ATP2A2	12	109207586	12q24.11	CARLA	+	build36
validation	ok				12	109207586			+	hg18
input data		rs3857067	N	SMARCAD1	4	95245457	4q22.2	CARLA	-	build36
validation	strand is incorrect				4	95245457			-	hg18
validation	strand flipped: - -> +				4	95245457			+	hg18
input data		rs4493911	N	LOXL2	8	23229408	8p21.3	CARLA	+	build36
validation	ok				8	23229408			+	hg18
input data		rs7765828	N	GMPR	6	16402701	6p22.3	CARLA	+	build36
validation	ok				6	16402701			+	hg18
input data		rs938291	N	SP3	2	174450854	2q31.1	CARLA	+	build36
validation	ok				2	174450854			+	hg18
input data		rs9920	N	CAV1	7	115987328	7q31.2	CARLA	+	build36
validation	ok				7	115987328			+	hg18

Output of „RepliCheckSNP“ - 1. checkPOS

strand_genotyped_SNP	NCBI_genome_build_used,_i.e._to_which_"chr"_and_"pos"_above_do_refer	strand_genotyped_SNP	target_sequ			.
strand_genome	**BUILD_used_NCBI_or_USCS**	**strand_genome**	**genomic_sequ_flanks**	**startPos sequ flanks**	**endPos sequ flanks**	**coded_allele_genotyped_SNP**
+	build36	+	AGCTTGCCTGGCCCTGAGCCTGAAGCGGCCTGAGAACCTGGTCTCTGTCCA			T
+	hg18	+	AGCTTGCCTGGCCCTGAGCCTGAAGCGGCCTGAGAACCTGGTCTCTGTCCA	61366301	61366351	C
+	build36	+	TTACACGTTAGCTCCTGGGAGGAGACAGGAGGGTGAAAACAACCTGGAGAC			T
+	hg18	+	TTACACGTTAGCTCCTGGGAGGAGATAGGAGGGTGAAAACAACCTGGAGAC	102044727	102044777	T
+	build36	+	TGCCATTTTCATTATATCCTCCTTCCTTCCCTGAACCAGTGTCCTGTCTTA			T
+	hg18	+	TGCCATTTTCATTATATCCTCCTTCTTTCCCTGAACCAGTGTCCTGTCTTA	200868919	200868969	T
+	build36	+	ATTCTCTTCTTGATTACCAACTTTGCTCTAAATGTAACACATCGTATGGTT			C
+	hg18	+	ATTCTCTTCTTGATTACCAACTTTGTTCTAAATGTAACACATCGTATGGTT	109207561	109207611	T
-	build36	-	TCTATTTTAATTATATGGAAAGGTAATTGCATCATCTCAATTAAGTTGTAT			T
-	hg18	-	ATACAACTTAATTGAGATGATGCAAATACCTTTCCATATAATTAAAATAGA	95245432	95245482	
+	hg18	+	TCTATTTTAATTATATGGAAAGGTATTTGCATCATCTCAATTAAGTTGTAT	95245432	95245482	T
+	build36	+	CCGACAGGGAAGTGGCTGTTCCTCTCTGCTGGGATGCTTTCCCCCGGGGAG			T
+	hg18	+	CCGACAGGGAAGTGGCTGTTCCTCTTTGCTGGGATGCTTTCCCCCGGGGAG	23229383	23229433	T
+	build36	+	AAGACACTTGATTTCTGATCTTAGACCAGACAGGCGGGAGGTGAAGCTCTG			G
+	hg18	+	AAGACACTTGATTTCTGATCTTAGACCAGACAGGCGGGAGGTGAAGCTCTG	16402676	16402726	C
+	build36	+	TTATCAACTGAAACTGAAATACCTACACTTTAACAAGATGTTAAGTATGTA			G
+	hg18	+	TTATCAACTGAAACTGAAATACCTAGACTTTAACAAGATGTTAAGTATGTA	174450829	174450879	G
+	build36	+	CTCCCTGAAGACCAAAATTAGAATACCCATGACCTAGTTTTCCATGCGTGT			C
+	hg18	+	CTCCCTGAAGACCAAAATTAGAATATCCATGACCTAGTTTTCCATGCGTGT	115987303	115987353	T

Flowchart of „RepliCheckSNP“ Software tool

A. Input:

1. One input file per study
2. File contains assay design and genotyping results information from each study

B. Web based query:

1. SNP flanking sequences from UCSC
2. SNP genotyping information from HapMart on chosen population(using wget)

C. Output - Three software routines generate three output files per study:

1. checkPOS:
   - Presence of SNP in chosen genome assembly (currently NCBI Build 37)
   - Correctness of given position to actual position of sequence in the genome
2. checkBLAT: Alignment of SNP probe sequence to chosen genome assembly
3. checkHWE:
   - Compares MAF and p(HWE) to HapMap using chosen population (e.g.CEU)
   - Detects strand (+/-) switches (e.g. C/T -> G/A) and coded allele switches (e.g. C/T -> T/C)

Output of „RepliCheckSNP“ - 2. checkBLAT

UCSC BLAT returns the output - which is formatted into an Excel or TXT spreadheet:

<table>
<tr><th>SOURCE</th><th>SNP_ID</th><th>LOCUS</th><th>GENO</th><th>CHR</th><th>POSITION</th><th>START</th><th>END</th><th></th><th></th><th>BUILD</th><th>STRAND</th><th>PROBE_sequence</th></tr>
<tr><th></th><th>BLAT_HIT#</th><th></th><th>TYPE</th><th>CHR</th><th></th><th>START</th><th>END</th><th>SCORE</th><th>%IDENT</th><th>DB_ID</th><th>STRAND</th><th>ALIGNED_sequences</th></tr>
<tr><td>input data</td><td>rs174583
...</td><td>FADS2</td><td>T/C</td><td>chr11</td><td>61366326</td><td>61366301</td><td>61366351</td><td></td><td></td><td>build36</td><td>+</td><td>AGCTTGCCTGGCCCTGAGCCTGAAG[C/T]GGCCTGAGAACCTGGTCTCTGTCCA</td></tr>
<tr><td>validation</td><td>Blat Hit#1
...
...
...</td><td></td><td></td><td>chr11</td><td></td><td>61366300</td><td>61366351</td><td>51.0</td><td>100</td><td>hg18</td><td>+</td><td>AGCTTGCCTGGCCCTGAGCCTGAAGCGGCCTGAGAACCTGGTCTCTGTCCA
||||||||||||||||||||||||||||||||||||||||||||||||||
AGCTTGCCTGGCCCTGAGCCTGAAGCGGCCTGAGAACCTGGTCTCTGTCCA</td></tr>
<tr><td>validation</td><td>Blat Hit#2
...
...
...</td><td></td><td></td><td>chr12</td><td></td><td>117871242</td><td>117871218</td><td>22.0</td><td>45.1</td><td>hg18</td><td>-</td><td>cctTTGtCcttggaTctGCaacctttGcCCTGAGAAcCtgGtctcCTGTCCA
||| | | || | |||||#| | | |||||||
AGCTTGCCTGGCCCTGAGCCTGAAGCGGCCTGAG_AACCTGGTCTCTGTCCA</td></tr>
<tr><td>input data</td><td>rs2273905
...</td><td>ANKRD9</td><td>T/C</td><td>chr14</td><td>102044752</td><td>102044727</td><td>102044777</td><td></td><td></td><td>build36</td><td>+</td><td>TTACACGTTAGCTCCTGGGAGGAGA[C/T]AGGAGGGTGAAAACAACCTGGAGAC</td></tr>
<tr><td>validation</td><td>Blat Hit#1
...
...
...</td><td></td><td></td><td>chr14</td><td></td><td>102044726</td><td>102044777</td><td>49.0</td><td>98</td><td>hg18</td><td>+</td><td>TTACACGTTAGCTCCTGGGAGGAGAtAGGAGGGTGAAAACAACCTGGAGAC
|||||||||||||||||||||||||| ||||||||||||||||||||||||||
TTACACGTTAGCTCCTGGGAGGAGACAGGAGGGTGAAAACAACCTGGAGAC</td></tr>
<tr><td>input data</td><td>rs295140
...</td><td>LOC26010</td><td>T/C</td><td>chr2</td><td>200868944</td><td>200868919</td><td>200868969</td><td></td><td></td><td>build36</td><td>+</td><td>TGCCATTTTCATTATATCCTCCTTC[C/T]TTCCCTGAACCAGTGTCCTGTCTTA</td></tr>
<tr><td>validation</td><td>Blat Hit#1
...
...
...</td><td></td><td></td><td>chr2</td><td></td><td>200868918</td><td>200868969</td><td>49.0</td><td>98</td><td>hg18</td><td>+</td><td>TGCCATTTTCATTATATCCTCCTTCtTTCCCTGAACCAGTGTCCTGTCTTA
|||||||||||||||||||||||||| ||||||||||||||||||||||||||
TGCCATTTTCATTATATCCTCCTTCCTTCCCTGAACCAGTGTCCTGTCTTA</td></tr>
<tr><td>validation</td><td>Blat Hit#2
...
...</td><td></td><td></td><td>chr4</td><td></td><td>170060451</td><td>170060410</td><td>24.0</td><td>52.9</td><td>hg18</td><td>-</td><td>TaagtagTcCAaagggTCCTCCtTcCtT_CCCTcActtttaaaCaaaaCcag
| | || ||||#| | | |#|||| | | |
TGCCATTTTCATTATATCCT_CCTTCCTTCCCTGAACCAGTGTCCTGTCTTA</td></tr>
</table>

Flowchart of „RepliCheckSNP“ Software tool

A. Input:

1. One input file per study
2. File contains assay design and genotyping results information from each study

B. Web based query:

1. SNP flanking sequences from UCSC
2. SNP genotyping information from HapMart on chosen population(using wget)

C. Output - Three software routines generate three output files per study:

1. checkPOS:
   - Presence of SNP in chosen genome assembly (currently NCBI Build 37)
   - Correctness of given position to actual position of sequence in the genome
2. checkBLAT: Alignment of SNP probe sequence to chosen genome assembly
3. checkHWE:
   - Compares MAF and p(HWE) to HapMap using chosen population (e.g.CEU)
   - Detects strand (+/-) switches (e.g. C/T -> G/A)
     and coded allele switches (e.g. C/T -> T/C)

Output of „RepliCheck“ Software - 3. checkHWE

SOURCE	Validation1	Validation2	Sentinel_SNP_ID_as_rs_number	SNP_replaced_by_proxy	most_plausible_gene_in_region	chromosome_number._Use_symbols_X,_XY,_Y_and_mt_for_non-autosomal_markers._	physical_position_for_the_reference_sequence	Band	study_in_which_SNP_was_typed	strand	coded_allele_genotyped_SNP	noncoded_or_reference_allele_genotyped_SNP	hapmap_allele_flip	strand_flip		coded_allele_frequ	noncoded_or_reference_allele_frequ	number_of_homozygotes_for_the_coded_allele	number_of_heterozygotes	number_of_homozygotes_for_the_noncoded_allele	n_total	chisq testvalue	HWE_pval	diff_frequ	chk_code	result
			Sentinel_SNP_ID_or_hapmap_dataset	SNP_proxy_necessary	LOCUS	CHR	POS		study_or_plattform																	
input data		.	rs174583	N	FADS2	11	61,366,326	11q12.2	CARLA	+	T	C			observed			188	684	669	1541	0.416	0.519			.
															expected	0.344	0.656	182.3	695.4	663.3						
hapmap		.	hm27_variation_ceu			11	61,366,326		AFFY_6.0	+	T	C			observed			15	49	49	113	0.243	0.622	0.006	0	.
															expected	0.350	0.650	13.8	51.4	47.8						
input data		HWE?	rs2273905	N	ANKRD9	14	102,044,752	14q32.31	CARLA	+	T	C			observed			185	744	615	1544	3.094	0.079			HWE?
															expected	0.361	0.639	200.9	712.1	630.9						
hapmap	allele flip	HWE?	hm27_variation_ceu			14	102,044,752		Illumina_1M	+	C	T	x		observed			60	38	15	113	4.550	0.033	-0.060	1	HWE?
															expected	0.699	0.301	55.2	47.5	10.2						
input data		.	rs295140	N	LOC26010	2	200,868,944	2q33.1	CARLA	+	T	C			observed			271	747	527	1545	0.050	0.822			.
															expected	0.417	0.583	268.9	751.3	524.9						
hapmap	allele flip	.	hm27_variation_ceu			2	200,868,944		Illumina_1M	+	C	T	x		observed			36	58	19	113	0.286	0.593	0.008	1	.
															expected	0.575	0.425	37.4	55.2	20.4						
input data		.	rs3026445	N	ATP2A2	12	109,207,586	12q24.11	CARLA	+	C	T			observed			184	713	610	1507	1.218	0.270			.
															expected	0.359	0.641	193.9	693.3	619.9						
hapmap		.	hm27_variation_ceu			12	109,207,586		Illumina_1M	+	C	T			observed			12	49	52	113	0.008	0.928	-0.036	0	.
															expected	0.323	0.677	11.8	49.4	51.8						
input data		HWE?	rs3857067	N	SMARCAD1	4	95,245,457	4q22.2	CARLA	+	T	A			observed			360	733	452	1545	3.524	0.061			HWE?
															expected	0.470	0.530	341.6	769.8	433.6						
hapmap	strand flip	.	hm27_variation_ceu			4	95,245,457		AFFY_6.0	-	A	T		x	observed			22	60	31	113	0.534	0.465	-0.010	2	.
															expected	0.460	0.540	23.9	56.1	32.9						
input data		.	rs4493911	N	LOXL2	8	23,229,408	8p21.3	CARLA	+	T	C			observed			52	454	1032	1538	0.057	0.812			.
															expected	0.181	0.819	50.6	456.8	1030.6						
hapmap	allele flip	.	hm27_variation_ceu			8	23,229,408		Perlegen	+	C	T	x		observed			30	20	1	51	1.291	0.256	0.034	1	.
															expected	0.784	0.216	31.4	17.3	2.4						

Output of „RepliCheck“ Software - 3. checkHWE

SOURCE	Validation1	Validation2	Sentinel_SNP_ID_as_rs_number / Sentinel_SNP_ID_or_hapmap_dataset	SNP_replaced_by_proxy / SNP_proxy_necessary	most_plausible_gene_in_region / LOCUS	chromosome_number._Use_symbols_X,_XY,_Y_and_mt_for_non-autosomal_markers._ / CHR	physical_position_for_the_reference_sequence / POS	Band	study_in_which_SNP_was_typed / study_or_plattform	strand	coded_allele_genotyped_SNP	noncoded_or_reference_allele_genotyped_SNP	hapmap_allele_flip	strand_flip
input data		.	rs174583	N	FADS2	11	61,366,326	11q12.2	CARLA	+	T	C		
hapmap		.	hm27_variation_ceu			11	61,366,326		AFFY_6.0	+	T	C		
input data		HWE?	rs2273905	N	ANKRD9	14	102,044,752	14q32.31	CARLA	+	T	C		
hapmap	allele flip	HWE?	hm27_variation_ceu			14	102,044,752		Illumina_1M	+	C	T	x	
input data		.	rs295140	N	LOC26010	2	200,868,944	2q33.1	CARLA	+	T	C		
hapmap	allele flip	.	hm27_variation_ceu			2	200,868,944		Illumina_1M	+	C	T	x	
input data		.	rs3026445	N	ATP2A2	12	109,207,586	12q24.11	CARLA	+	C	T		
hapmap		.	hm27_variation_ceu			12	109,207,586		Illumina_1M	+	C	T		
input data		HWE?	rs3857067	N	SMARCAD1	4	95,245,457	4q22.2	CARLA	+	T	A		
hapmap	strand flip	.	hm27_variation_ceu			4	95,245,457		AFFY_6.0	-	A	T		x
input data		.	rs4493911	N	LOXL2	8	23,229,408	8p21.3	CARLA	+	T	C		
hapmap	allele flip	.	hm27_variation_ceu			8	23,229,408		Perlegen	+	C	T	x	

Output of „RepliCheck“ Software - 3. checkHWE

coded_allele_genotyped_SNP	noncoded_or_reference_allele_genotyped_SNP	hapmap_allele_flip	strand_flip		coded_allele_frequ	noncoded_or_reference_allele_frequ	n_homozygous_coded_allele	n_heterozygous	n_homozygous_noncoded_allele	n_total	chisq testvalue	HWE_pval	diff_frequ	chk_code	result
T	C			observed			188	684	669	1541	0.416	0.519			.
				expected	0.344	0.656	182.3	695.4	663.3						
T	C			observed			15	49	49	113	0.243	0.622	0.006	0	.
				expected	0.350	0.650	13.8	51.4	47.8						
T	C			observed			185	744	615	1544	3.094	0.079			HWE?
				expected	0.361	0.639	200.9	712.1	630.9						
C	T	x		observed			60	38	15	113	4.550	0.033	-0.060	1	HWE?
				expected	0.699	0.301	55.2	47.5	10.2						
T	C			observed			271	747	527	1545	0.050	0.822			.
				expected	0.417	0.583	268.9	751.3	524.9						
C	T	x		observed			36	58	19	113	0.286	0.593	0.008	1	.
				expected	0.575	0.425	37.4	55.2	20.4						
C	T			observed			184	713	610	1507	1.218	0.270			.
				expected	0.359	0.641	193.9	693.3	619.9						
C	T			observed			12	49	52	113	0.008	0.928	-0.036	0	.
				expected	0.323	0.677	11.8	49.4	51.8						
T	A			observed			360	733	452	1545	3.524	0.061			HWE?
				expected	0.470	0.530	341.6	769.8	433.6						
A	T		x	observed			22	60	31	113	0.534	0.465	-0.010	2	.
				expected	0.460	0.540	23.9	56.1	32.9						
T	C			observed			52	454	1032	1538	0.057	0.812			.
				expected	0.181	0.819	50.6	456.8	1030.6						
C	T	x		observed			30	20	1	51	1.291	0.256	0.034	1	.
				expected	0.784	0.216	31.4	17.3	2.4						

DDD - Design Development Dissemination of „RepliCheckSNP“

A. Project Collaborators

- Arne Pfeufer (design of task, development of Pflichtenheft)
- Dieter Amilo (coding of Software solution „RepliCheck)
- Christopher Newton-Cheh (valuable discussion of projects goals within QT-IGC)

B. RepliCheckSNP – Software distribution

- RepliCheckSNP is free of charge (publicly funded development)
- RepliCheckSNP will be downloadable in near future from
  - the homepage of the TMF e.V. (www-tmf-ev.de)
  - the homepage of the developer (www.helmholtz.muenchen.de/ihg)
- RepliCheckSNP download includes executable software (.jar) and sample files

DDD - Design Development Dissemination of „RepliCheckSNP“










universitätbonn

UKSH

MRI

TMF

dkfz. GERMAN CANCER RESEARCH CENTER

MAX-PLANCK-GESELLSCHAFT