Second Multi-IC Symposium
Working Group 2:

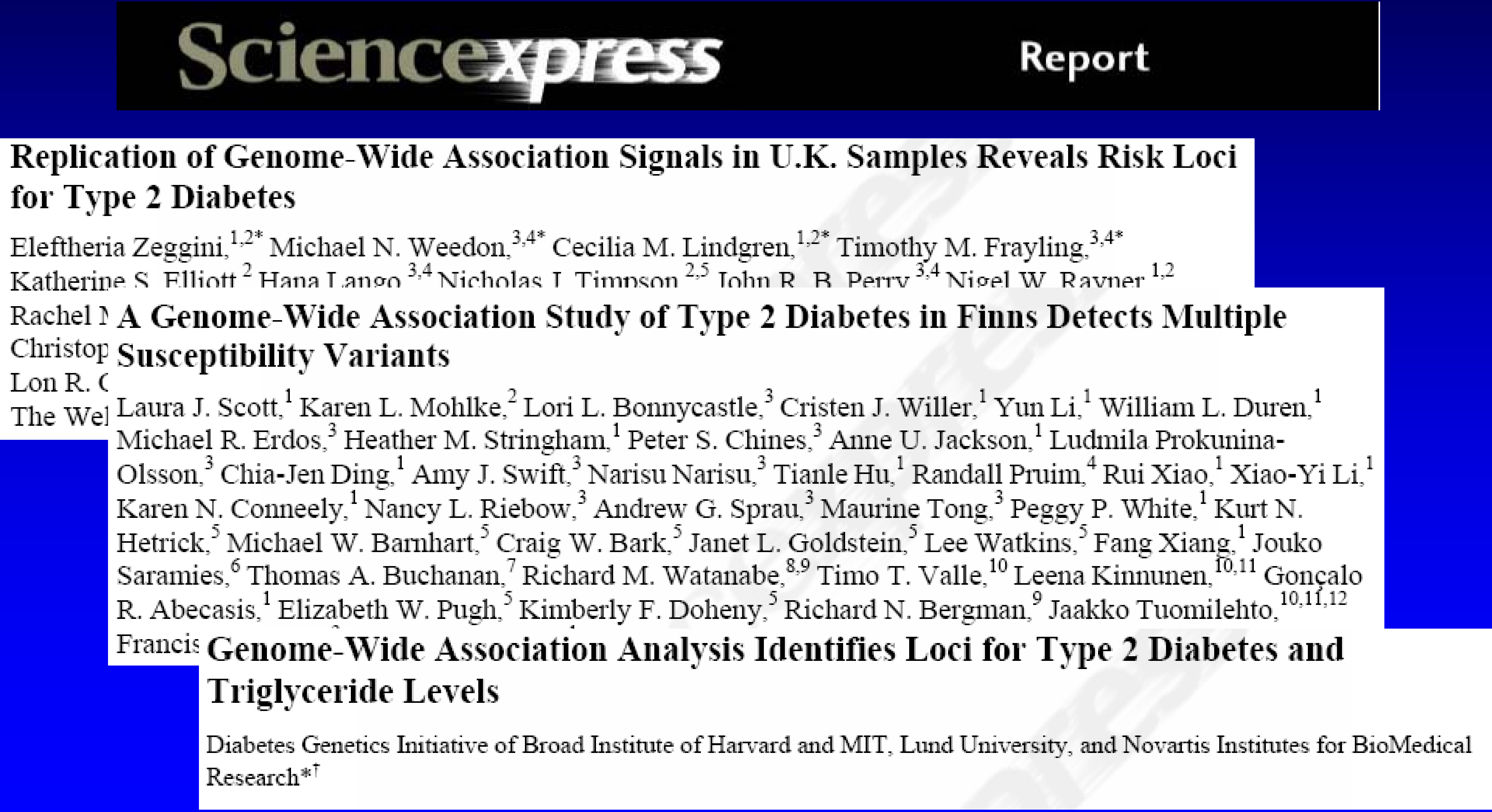
Facilitating Cross Study GWAS Analyses

Francis Collins, NHGRI
Richard Hayes, NCI
Catherine McKeon, NIDDK
Chris O’Donnell, NHLBI
Steve Sherry, NCBI

Facilitating Cross Study GWAS Analyses

- Strategies for the design, analysis, and reporting of results from such analyses.
- What are the best analysis strategies for combining different genotyping platforms?
- Assessing homogeneity or heterogeneity of cohort populations and phenotypes.
- Cross-study GWA involving multiple traits in two or more population-based cohorts.
- Using pools of GWA cohort(s) as a common set of GWA controls in case-control studies.
- How to foster inter-IC and international consortia and collaborations for such studies.

Diabetes Mellitus GWAS



Sciencexpress Report

Replication of Genome-Wide Association Signals in U.K. Samples Reveals Risk Loci for Type 2 Diabetes

Eleftheria Zeggini,[1,2]* Michael N. Weedon,[3,4]* Cecilia M. Lindgren,[1,2]* Timothy M. Frayling,[3,4]* Katherine S. Elliott,[2] Hana Lango,[3,4] Nicholas J. Timpson,[2,5] John R. B. Perry,[3,4] Nigel W. Rayner,[1,2]

Rachel M
Christop
Lon R. C
The Wel

A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants

Laura J. Scott,[1] Karen L. Mohlke,[2] Lori L. Bonnycastle,[3] Cristen J. Willer,[1] Yun Li,[1] William L. Duren,[1] Michael R. Erdos,[3] Heather M. Stringham,[1] Peter S. Chines,[3] Anne U. Jackson,[1] Ludmila Prokunina-Olsson,[3] Chia-Jen Ding,[1] Amy J. Swift,[3] Narisu Narisu,[3] Tianle Hu,[1] Randall Pruim,[4] Rui Xiao,[1] Xiao-Yi Li,[1] Karen N. Conneely,[1] Nancy L. Riebow,[3] Andrew G. Sprau,[3] Maurine Tong,[3] Peggy P. White,[1] Kurt N. Hetrick,[5] Michael W. Barnhart,[5] Craig W. Bark,[5] Janet L. Goldstein,[5] Lee Watkins,[5] Fang Xiang,[1] Jouko Saramies,[6] Thomas A. Buchanan,[7] Richard M. Watanabe,[8,9] Timo T. Valle,[10] Leena Kinnunen,[10,11] Gonçalo R. Abecasis,[1] Elizabeth W. Pugh,[5] Kimberly F. Doheny,[5] Richard N. Bergman,[9] Jaakko Tuomilehto,[10,11,12]

Francis

Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels

Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes for BioMedical Research*†

Three Groups Working Together Greatly Adds to Power

FUSION

S1: 1161 + 1174

S2: 1215 + 1258

DGI

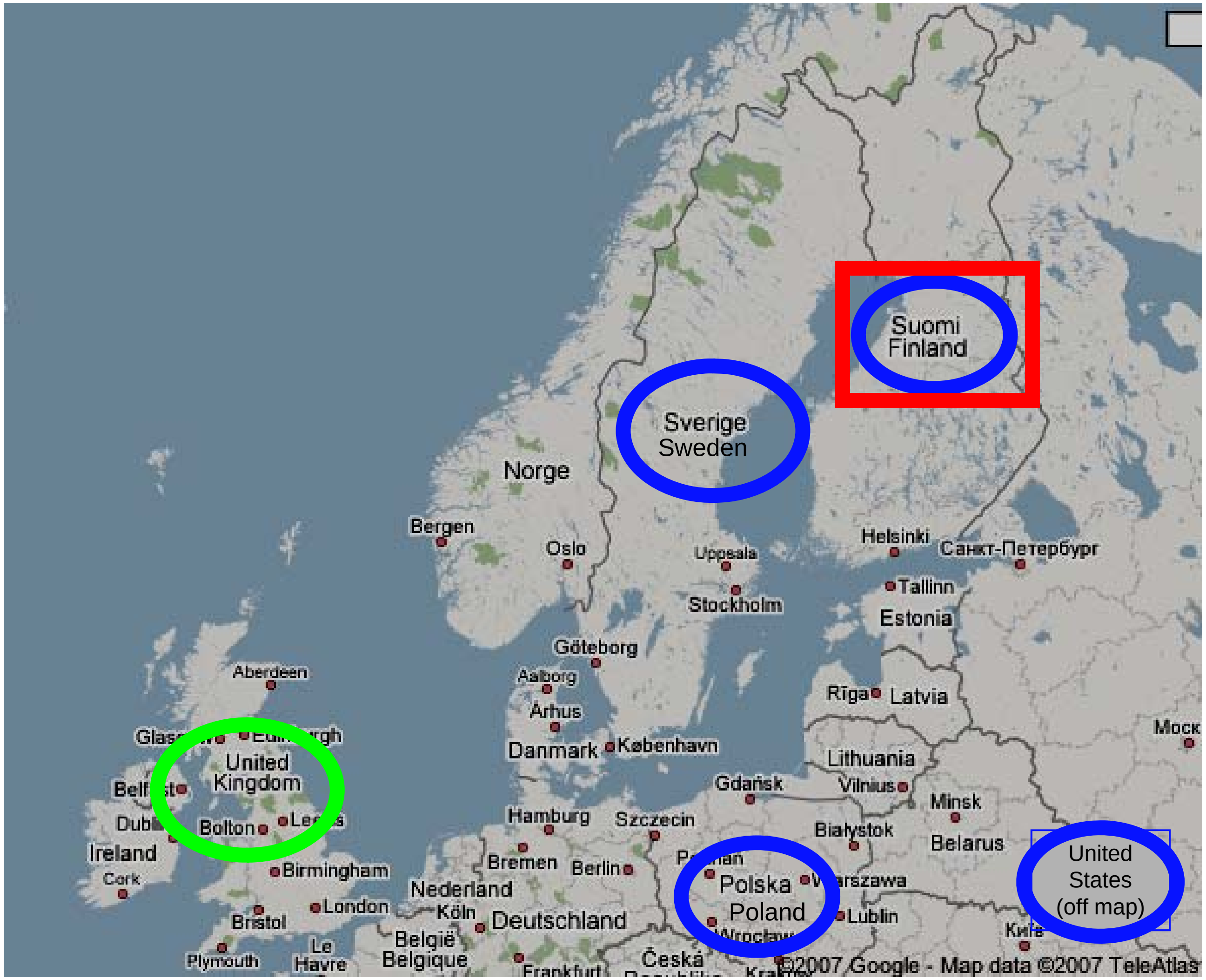
S1: 1464 + 1467

S2: 5065 + 5785

WTCCC/UKT2D

S1: 1924 + 2938

S2: 3757 + 5346

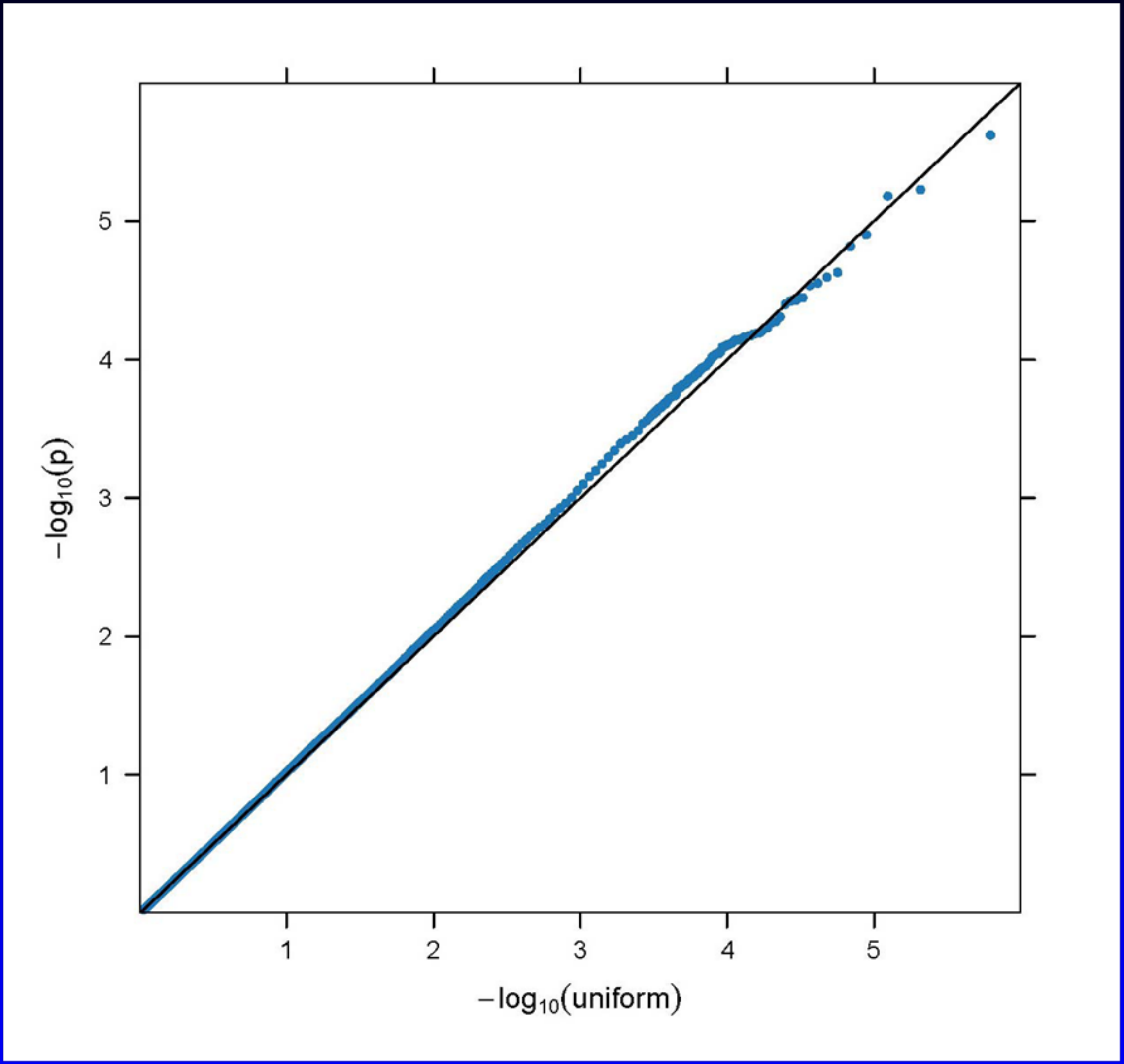


Totals

S1 = 4549 + 5579

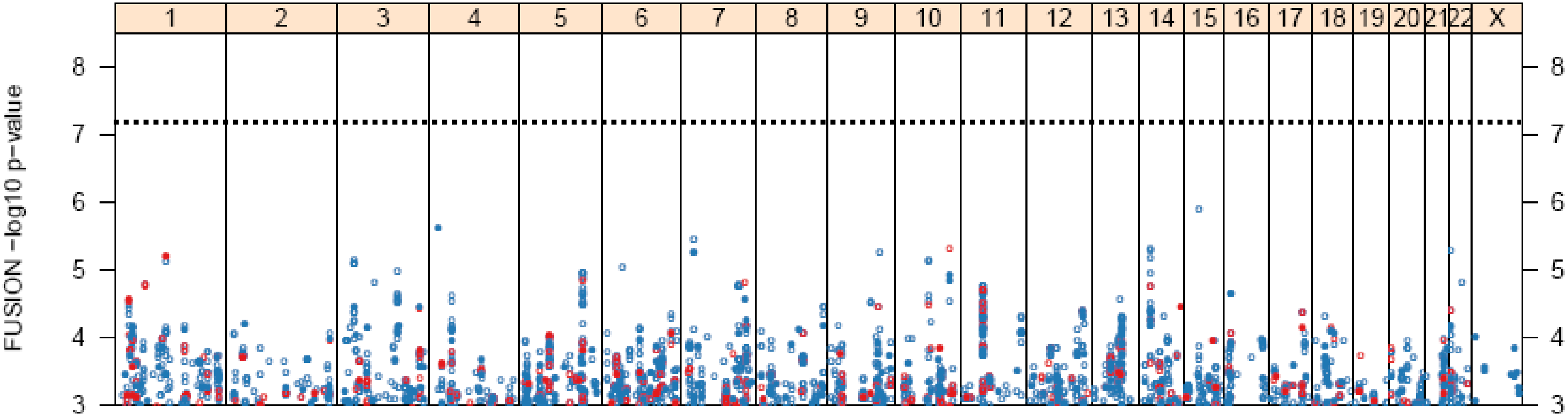
S2 = 10053 + 12389

(n=32,554)

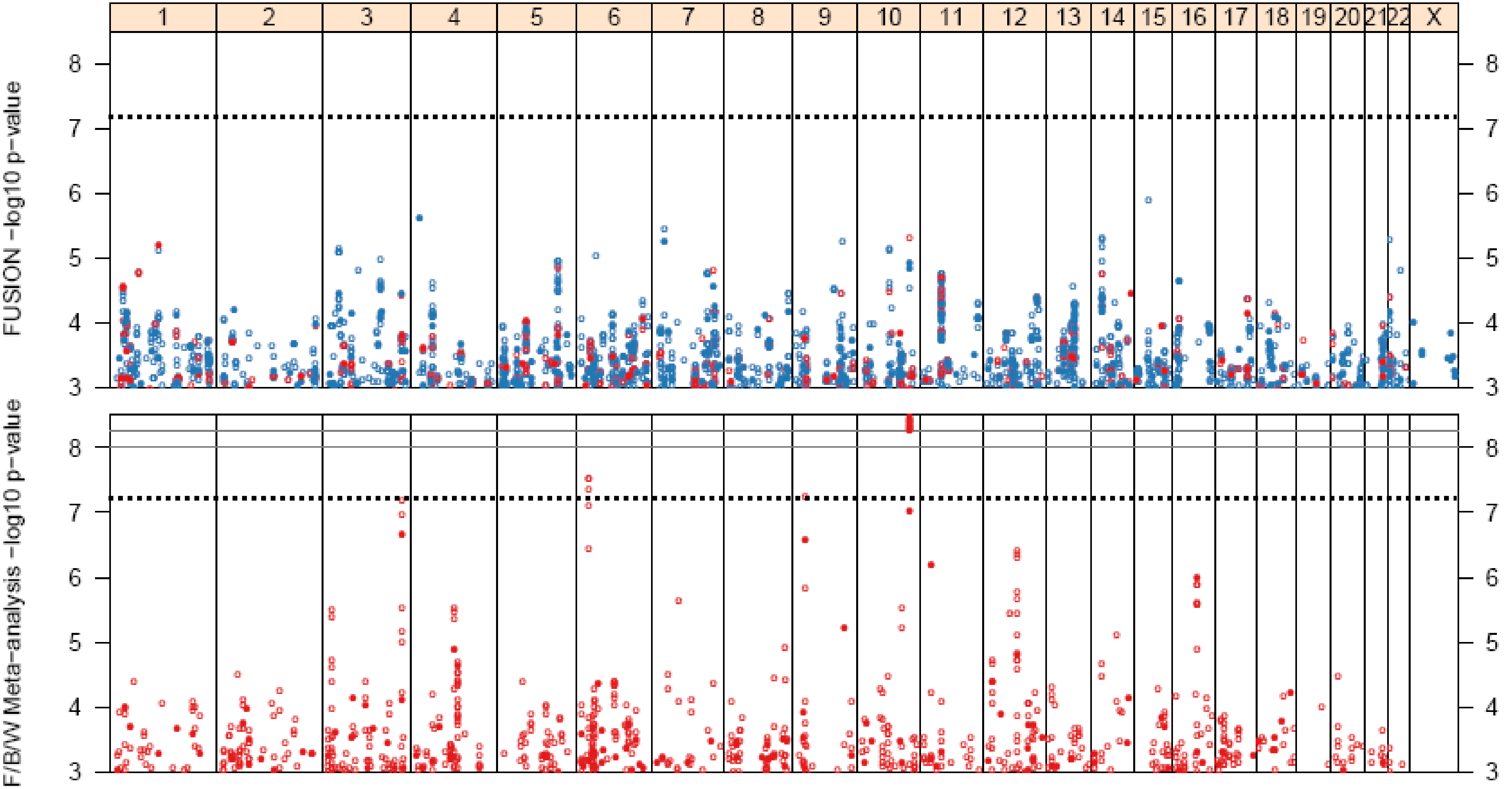


GOOD NEWS AND BAD NEWS:
Q-Q PLOT FOR FUSION SHOWS NO EVIDENCE FOR STRATIFICATION,
BUT NOT MUCH EVIDENCE FOR SUSCEPTIBILITY VARIANTS EITHER!

Stage 1: FUSION only (1161 cases + 1174 controls)



Stage 1 – FUSION only



Stage 1 – FUSION + DGI + WTCCC

(4549 cases + 5579 controls)

Imputing Missing Genotypes in Case Control Samples

- Methods and software have now been developed and tested by
  - Goncalo Abecasis, Michigan
  - Jonathan Marchini, Oxford
- Begins with GWA data from panel of choice
- Uses HapMap data from similar geographic origins to infer what alleles were most likely present at untyped loci
- Limited to SNPs in strong LD with typed SNPs
- Can produce quality score estimates
- Allows merging of data sets from Illumina, Affymetrix, or Perlegen panels

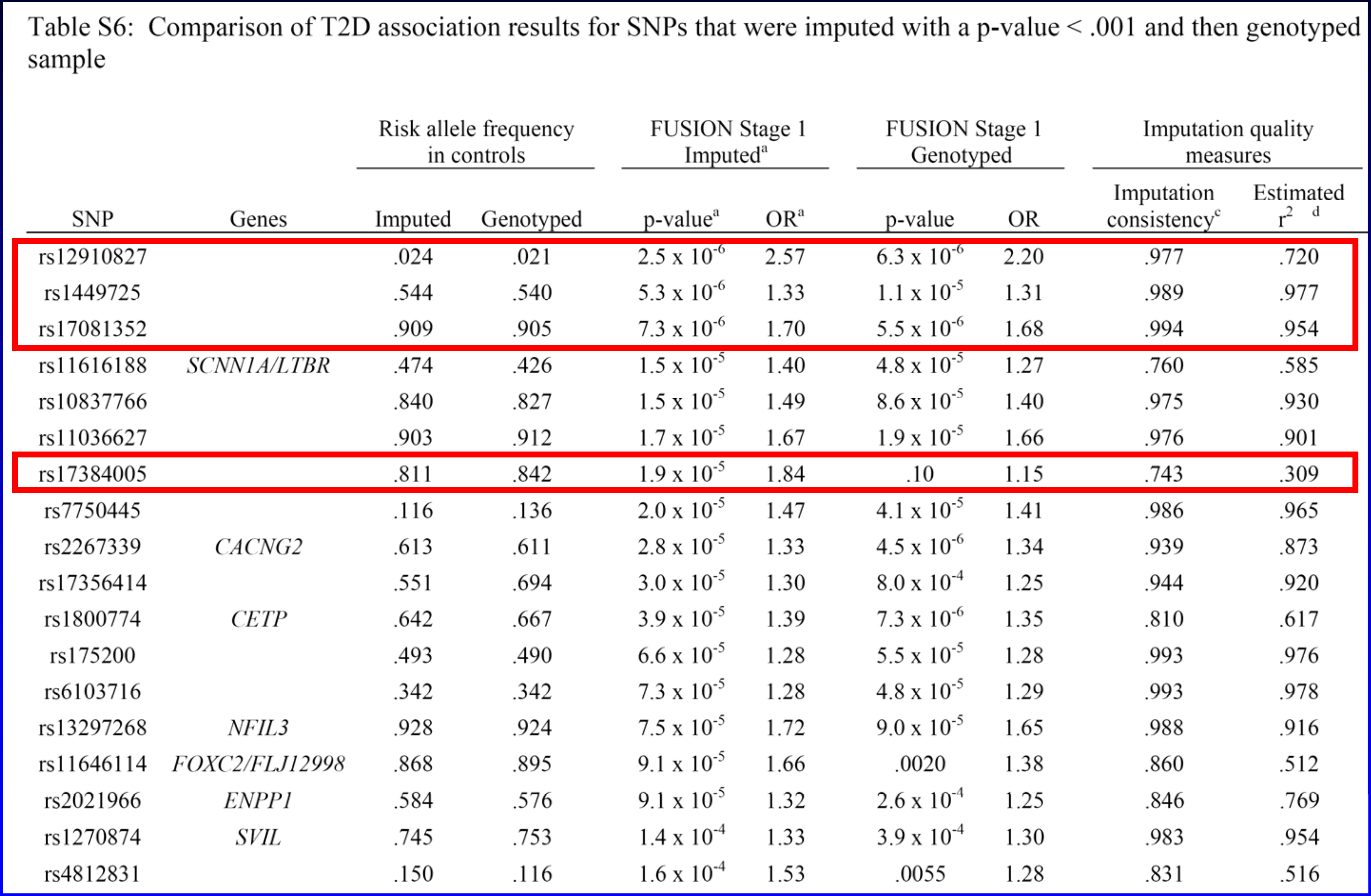


Table S6: Comparison of T2D association results for SNPs that were imputed with a p-value < .001 and then genotyped sample

		Risk allele frequency in controls		FUSION Stage 1 Imputed[a]		FUSION Stage 1 Genotyped		Imputation quality measures	
SNP	Genes	Imputed	Genotyped	p-value[a]	OR[a]	p-value	OR	Imputation consistency[c]	Estimated r^2 [d]
rs12910827		.024	.021	2.5 x 10^{-6}	2.57	6.3 x 10^{-6}	2.20	.977	.720
rs1449725		.544	.540	5.3 x 10^{-6}	1.33	1.1 x 10^{-5}	1.31	.989	.977
rs17081352		.909	.905	7.3 x 10^{-6}	1.70	5.5 x 10^{-6}	1.68	.994	.954
rs11616188	*SCNN1A/LTBR*	.474	.426	1.5 x 10^{-5}	1.40	4.8 x 10^{-5}	1.27	.760	.585
rs10837766		.840	.827	1.5 x 10^{-5}	1.49	8.6 x 10^{-5}	1.40	.975	.930
rs11036627		.903	.912	1.7 x 10^{-5}	1.67	1.9 x 10^{-5}	1.66	.976	.901
rs17384005		.811	.842	1.9 x 10^{-5}	1.84	.10	1.15	.743	.309
rs7750445		.116	.136	2.0 x 10^{-5}	1.47	4.1 x 10^{-5}	1.41	.986	.965
rs2267339	*CACNG2*	.613	.611	2.8 x 10^{-5}	1.33	4.5 x 10^{-6}	1.34	.939	.873
rs17356414		.551	.694	3.0 x 10^{-5}	1.30	8.0 x 10^{-4}	1.25	.944	.920
rs1800774	*CETP*	.642	.667	3.9 x 10^{-5}	1.39	7.3 x 10^{-6}	1.35	.810	.617
rs175200		.493	.490	6.6 x 10^{-5}	1.28	5.5 x 10^{-5}	1.28	.993	.976
rs6103716		.342	.342	7.3 x 10^{-5}	1.28	4.8 x 10^{-5}	1.29	.993	.978
rs13297268	*NFIL3*	.928	.924	7.5 x 10^{-5}	1.72	9.0 x 10^{-5}	1.65	.988	.916
rs11646114	*FOXC2/FLJ12998*	.868	.895	9.1 x 10^{-5}	1.66	.0020	1.38	.860	.512
rs2021966	*ENPP1*	.584	.576	9.1 x 10^{-5}	1.32	2.6 x 10^{-4}	1.25	.846	.769
rs1270874	*SVIL*	.745	.753	1.4 x 10^{-4}	1.33	3.9 x 10^{-4}	1.30	.983	.954
rs4812831		.150	.116	1.6 x 10^{-4}	1.53	.0055	1.28	.831	.516

Top 10 Results From Combined Analysis Of Stage 1 + Stage 2 From All Three Groups

14602 cases + 17968 controls

	FUSION		DGI		WTCCC/UKT2D		All Samples	
Gene	OR	p-value	OR	p-value	OR	p-value	OR	p-value
TCF7L2	1.34	1.3×10^{-8}	1.38	2.3×10^{-31}	1.37	6.7×10^{-13}	1.37	1.0×10^{-48}
IGF2BP2	1.18	2.1×10^{-4}	1.17	1.7×10^{-9}	1.11	1.6×10^{-4}	1.14	8.9×10^{-16}
CDKN2A/B	1.20	.0022	1.20	5.4×10^{-8}	1.19	4.9×10^{-7}	1.20	7.8×10^{-15}
FTO	1.11	0.016	1.03	0.25	1.23	7.3×10^{-14}	1.17	1.3×10^{-12}
CDKAL1	1.12	0.0095	1.08	0.0024	1.16	1.3×10^{-8}	1.12	4.1×10^{-11}
KCNJ11	1.11	0.013	1.15	1.0×10^{-7}	1.15	0.0013	1.14	6.7×10^{-11}
HHEX	1.10	0.026	1.14	1.7×10^{-4}	1.13	4.6×10^{-6}	1.13	5.7×10^{-10}
SLC30A8	1.18	7.0×10^{-5}	1.07	0.047	1.12	7.0×10^{-5}	1.12	5.3×10^{-8}
Chr 11	1.48	5.7×10^{-8}	1.16	0.12	1.13	0.068	1.23	4.3×10^{-7}
PPARG	1.20	0.0014	1.09	0.019	1.23	0.0013	1.14	1.7×10^{-6}

Strategies for the design, analysis, and reporting of results from such analyses.

- Advance planning for in silico comparisons:
  - Selection of similarly defined phenotype(s)
  - Conduct of similar covariate adjustment
  - Criteria for QC and genotype filtering criteria
- Should data be compared at level of individual participant data or aggregate GWA results?
- Who conducts the analysis?
- Publication strategies: options for assigning authorship and writing publications? How can junior investigators play a key authorship role?
- What if data sharing policies differ?
- Merits and drawbacks of rapid web-posting of in silico comparison results?

Best analysis strategies for different genotyping platforms?

- Imputation of genotypes using HapMap
- How are these analyses conducted?
- What are the best algorithms available?
- What are the controversies about the available algorithms?
- What role can be played by dbGaP?
- What other genetic variation be captured by the available techniques (copy number variation, rare sequence variants)?

dbGaP plan for the distribution of imputed genotype data

- Original data sets are clearly labeled by study accession (phs#) and analysis version (phg#).
- Imputed genotypes distributed separately from original data with clear attribution of method, estimated quality and scope (with consent of PI).
- 2 imputation activities
  - Replacing missing data within a platform
  - Estimating additional untyped markers for cross-platform comparisons

Assessing homogeneity or heterogeneity of cohort populations and phenotypes.

- Disease-based case-control or case-cohort versus prospective observational cohorts
- Quantitative vs dichotomous/disease traits
- Phenotype definition; sources of heterogeneity
- Use of covariate-adjusted phenotypes
- Assessment for modification by age and sex
- GWAS studies in populations of different ethnicities
- When to test for population stratification

Cross-study GWA involving multiple traits in two or more population-based cohorts.

- Identifying planned or ongoing GWAS in population-based cohorts
- Identifying and accessing phenotypes in cohorts with GWAS, e.g.
  - GAIN
  - WTCCC
  - NHLBI SHARe, CARE and STAMPEED
  - NCBI CGEMs
  - GEI
- Logistical challenges to inter-cohort studies:
  - Single investigators vs central Steering Committee
  - Differences in publication and sharing strategies
  - Differences in informed consent

Using pools of GWA cohort(s) as a common set of GWA controls in case-control studies.

- Pros: Increased sample size, ability to study relevant subgroups (e.g., age, sex, cig smokes)
- Cons: Population heterogeneity, clear documentation of “control” (i.e., absence of case status) may be absent
- Identifying and accessing GWAS data sources amenable to such approaches
  - GAIN
  - WTCCC
  - NHLBI SHARe, CARE and STAMPEED
  - NCBI CGEMs
- dbGaP “universal controls” currently being submitted by Illumina and GSK in addition to study-specific control datasets.

How to foster inter-IC and international consortia and collaborations for such studies.

- Inter-IC consortia and collaborations
  - Disease-based (e.g., Diabetes, Cancer)
  - Cohort-based (e.g., NHLBI cohorts)
  - Pathophysiology-based (e.g., Inflammation)
  - Systems biology-based
- Can we look beyond disease-based silos?
- International consortia:
  - Examples: WTCC, German National Genome Research Network
  - Challenges, opportunities
  - Handling differences in data sharing policies