GPU enables routine bivariate & trivariate analysis of Case-Control GWAS

*Qiao Wang, Fan Shi, Andrew Kowalczyk, David Rawlinson, Justin Bedo, Cheng Soon Ong, Benjamin Goudey, Richard Campbell, Herman Ferra and **Adam Kowalczyk***

National ICT Australia, Victorian Research Laboratory
&
Centre for Neural Engineering
The University of Melbourne

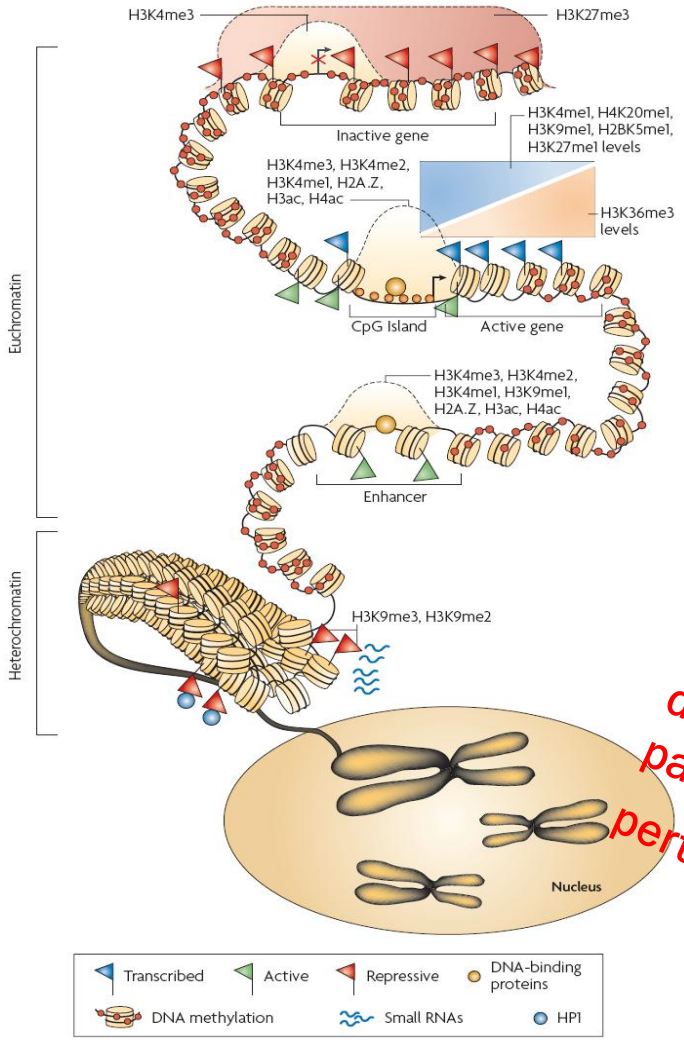
kowa@unimelb.edu.au
http://bioinformatics.research.nicta.com.au

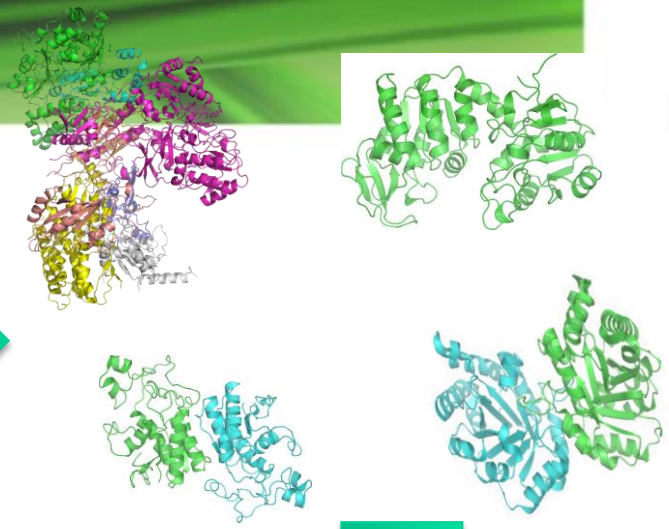
GPU Technology Conference 25/03/2014

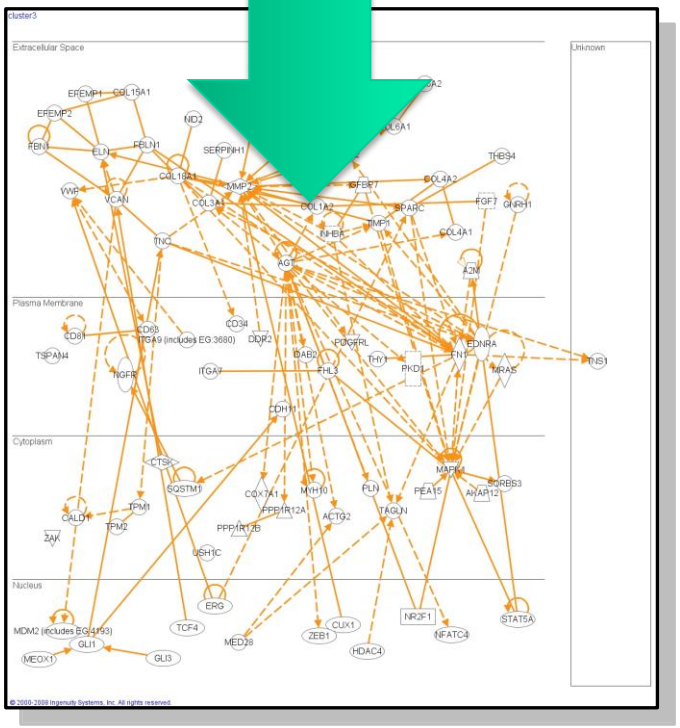
Overview

- The problem
- GPU solution
- Sample Results
- Way forward

GWAS Expectations





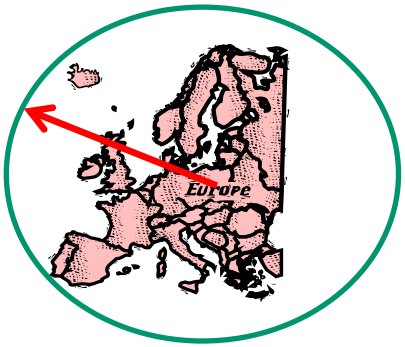


High throughput detection of pathway perturbation?

Dustin E. Schones and Keji Zhao,, Genome-wide approaches to studying chromatin modifications, nature reviews, **genetics** volume 9 , march

GWAS challenge and Promise

Human genome
= 2 x 3,000 km!



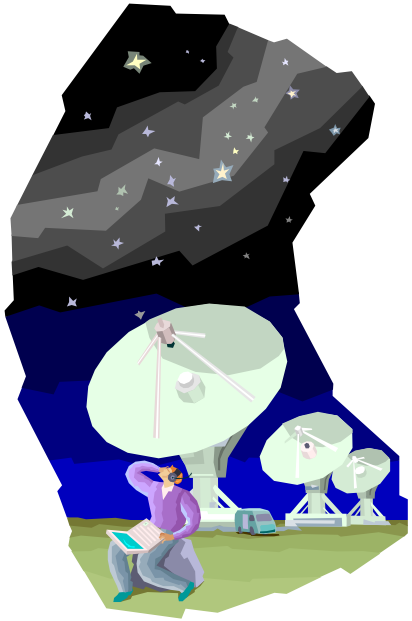
Genome-wide association studies

- 1500 studies, ~$1 B
- 500K ➔ 5M probes ➔ ?? NGS
- 1000 ➔ 100,000 subjects

Personalized medicine
- treatment selection
- detection of side effect risk

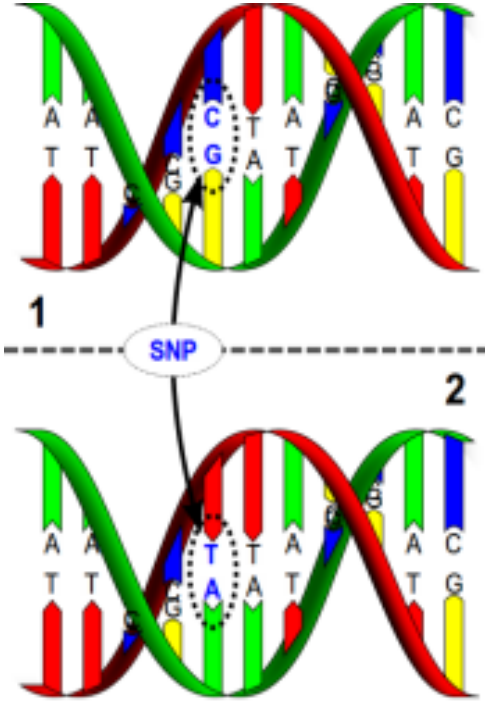
Direct genotype services (e.g. 23andme.com)
- life style managements if disease risk

Pharma
- drug target identification
- drug repositioning



Genotyping Markers

- SNPs: Single or Simple Nucleotide Polymorphism
  - 1 base substitution or
  - few bases insertion / deletion
- Human Genome: 3 billion base pairs
  - ~ 10 million SNPs



- Aim: find SNPs that are associated with given trait / risk of disease
- Genome Wide Association Studies (GWAS):
  genotype SNPs across the whole genomes of multiple individuals

Genome-Wide Association Studies (GWAS) Data

NICTA

CASE

CONTROL

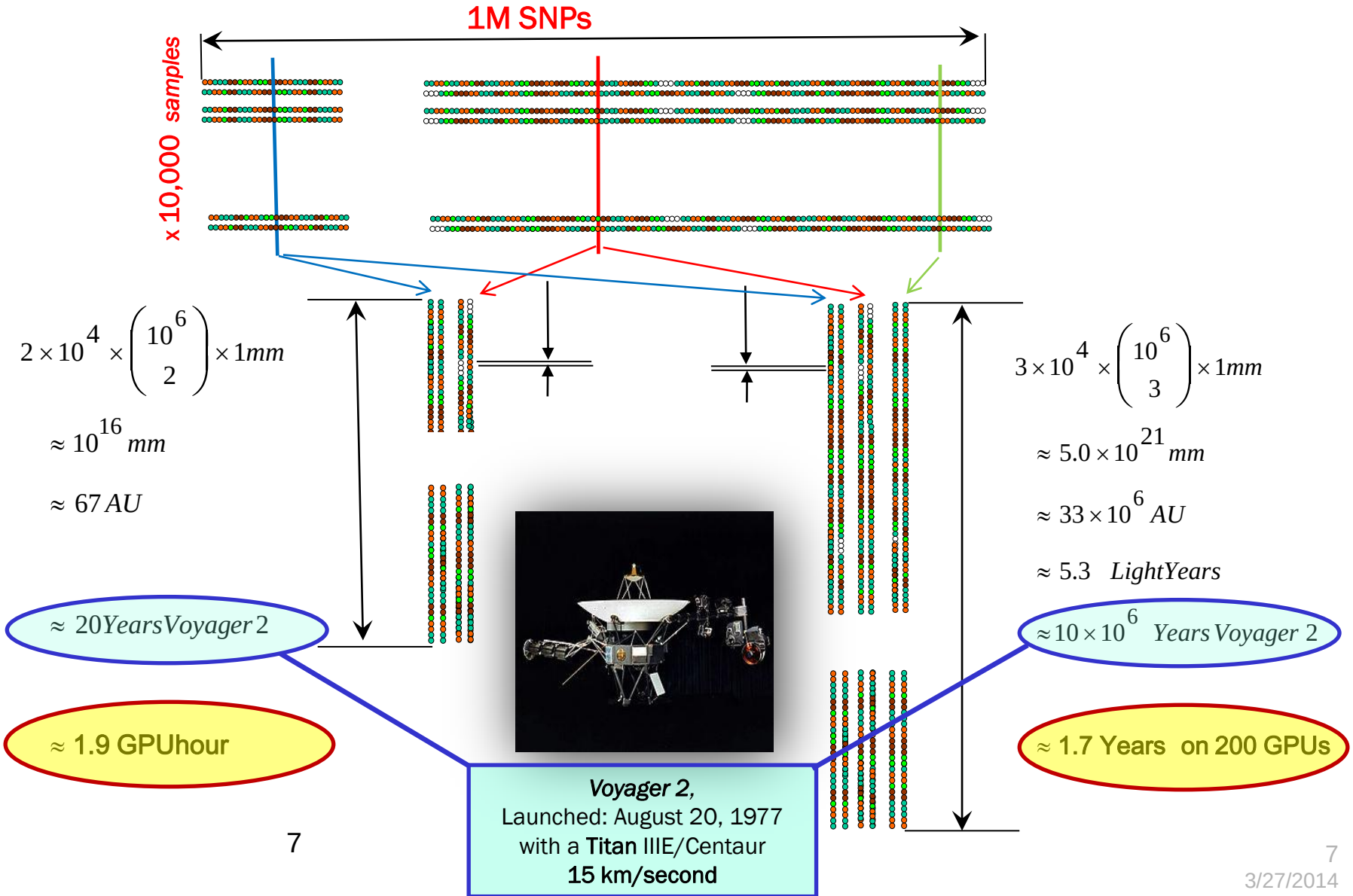
	SNP A	SNP B	...	...	SNP *N*
Person 1	1	2	0	2	1
Person 2	0	-	1	0	2
...	2	1	0	1	1
...	1	-	2	2	2
...	2	2	1	2	0
Person *M*	0	0	2	0	1

***M* = 10 K**

***N* = 1 M**

0 (AA) = dominant homozygous, 1 (Aa) = heterozygous, 2 (aa) = recessive homozygous

Size of Epistasis Search Space



Three challenges for epistasis detection

Computational:

How to examine trillions of candidates ?

Algorithmic:

How to detect potential epistasis efficiently?

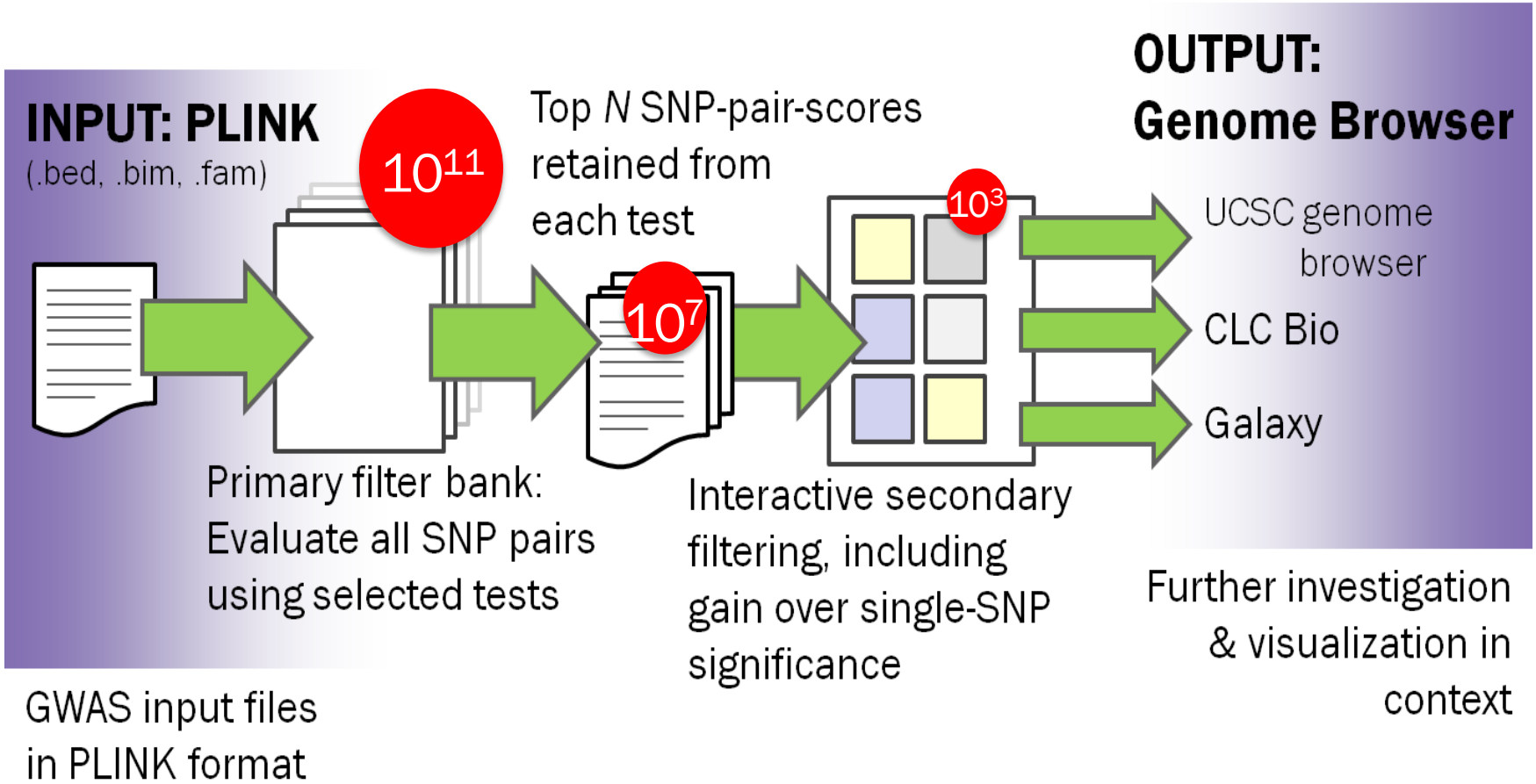
Statistical:

How to combat noise injected by multiple choices selection?

GWIS: Genome Wide Interaction Search
(GPU & CPU variants)

Need to reduce candidate interactions to small number, for in-depth analysis by biologists

A screening problem



GalaxyProject Web-Server instance

We are now hosting a GalaxyProject web server with **GPU** engine allows use of GWIS

- Free for non-commercial use
- Exhaustive bivariate analysis
- Hardware = Single desktop computer with standard GPU
- Input data in PLINK format

http://gwis1.research.nicta.com.au/

Timing at a Glance

Desktop computer

SNPs x Samples	Standard algorithm (IG*)	GWIS-CPU (4 Cores Intel 3.0 GHz)	GWIS-GPU (1 GTX 470) *Chi-squared test*	GWIS-GPU on MASSIVE GPU Cluster (~ 200 Tesla C2050)	GWIS-GPU on Titan (18,688 Tesla K20)
2nd Order					
300K x 3K	108 days	39 min	3 min	~ 0	~ 0
1M x 10K	11 years	25 hours	1.85 hours	~ 0.5 min.	~ 0
5M x 10K	275 years	26 days	1.91 days	~ 12.24 min.	~ 0
3rd Order					
300K x 3K	~ 30K years	~ 30 years	~ 2.3 years	~ 5 days	~ 38 min
1M x 10K	~ 3.6M years	~ 3.7K years	~ 282 years	~ 612 days	~ 3.2 days
5M x 10K	~ 458.3M years	~ 453K years	~ 34.9K years	~ 208 years	~ 1.1 years

Feasible with moderate $$

Approaching feasibility

$GWIS_{FI}$ Timing Benchmark

$GWIS_{FI}$ ⇔ GWIS functional interface

an easy to program library of dedicated kernels (Wang *et.al*, 2014)

	Methods	Original algorithm: Hardware Platform	450K x 5K	5M x 10K	$GWIS_{FI}$ - **GPU** implementation: 450K x 5K	5M x 10K	Time Improvement
		GPU implementations (*all timings, measured or scaled*, are for 1x Nv GTX 470 (1.088 Tflops)*)			**Nv GTX 470** (1.088 Tflops) (*timings for 1x Nv GTX 470*)		
1	GBOOST	*1x Nv GTX 470*	1.4 hours	14.4* days	14 mins	2.4* days	**6×**
2	SHEsisEpi♭	**2x** Nv GTX 285 (1.062 Tflops)	21.8* hours	7.5* months	12 mins	2.1* days	**109×**
3	EPIBLASTER♮ (stage 1)	**4x** Nv GTX 295 (1.789 Tflops)	48.6* hours	1.37* years	9 mins	1.5* days	**324×**
4	GWIS1: SS, DSS & χ^2 *together*	*1x Nv GTX 470*	23 mins	3.95* days	20 mins	3.4* days	**1.15×**
		CPU implementations			**Nv GTX 470** (1.088 Tflops) (*timings for 1x Nv GTX 470*)		
5	FastEpistasis (PLINK 1.9)	*Intel Core i7 3.07GHz*	9.0 hours	3.1* months	10 mins	1.7* days	**54×**
6	FastEpistasis (PLINK 1.07)	*Intel Core i7 3.07GHz*	2.8* months	57.6* year	10 mins	1.7* days	**> 12,000×**
7	IG‡	1 CPU, 3 GHz	3.7* years	907* years	10 mins	1.7* days	**$\gg 10^5$×**
8	$2\chi^2$ §	1 CPU, 3 GHz	18* days	12* years	10 mins	1.7* days	**$\gg 10^5$×**
					Nv GTX Titan (4.5 Tflops) (*timings for Nv GTX Titan*)		
9	**All above methods together**				**18 mins**	**3.1* days**	-

Sample results

What is epistasis?

- Multiple definitions

- Our working definition:

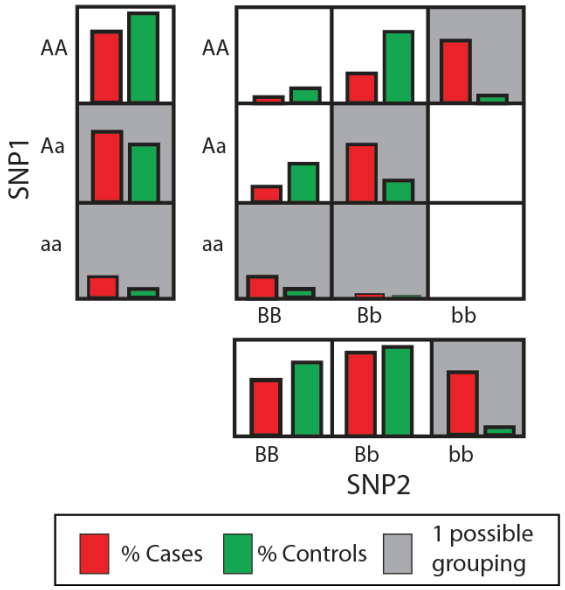
Two SNPs have an epistatic interaction if using both of them allows discrimination between Cases and Controls with ***significantly*** *higher sensitivity and specificity than is possible using any one of them individually.*

See: B. Goudey, D. Rawlinson, Q. Wang, F. Shi, H. Ferra, R. M. Campbell, L. Stern, M. T. Inouye, C. Soon Ong, A. Kowalczyk, ***GWIS - Model-free, Fast and Exhaustive Search for Epistatic Interactions in Case-Control GWAS****, BMC Genomics 2013*

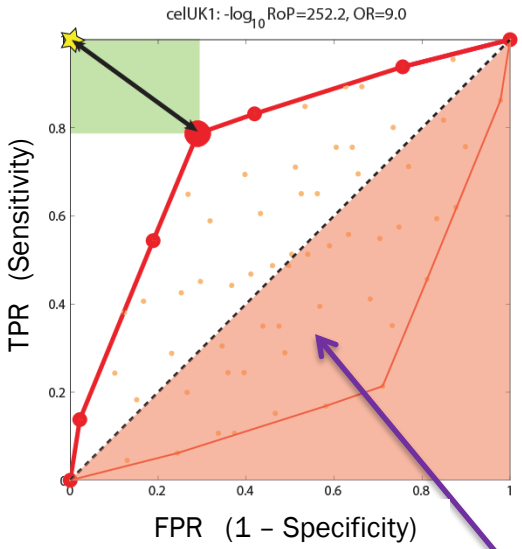
The optimal ROC curves based significance tests for epistasis

For any observed sensitivity and specificity achieved by a pair of SNP-probes, compute the probability of observing equal or higher specificity and sensitivity due to **biased sampling** from the population for which true specificity and sensitivity falls in the null hypothesis region

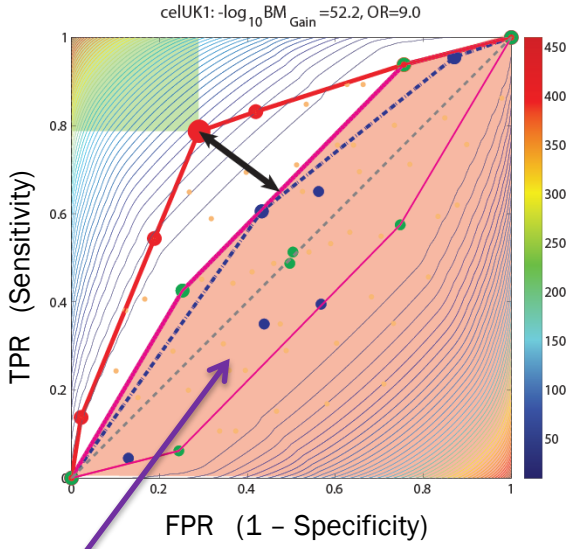
Classifier construction from contingency table



Sensitivity-Specificity test



Gain in Sensitivity-Specificity test (GSS)



snp$_a$ snp$_b$ convHull snp$_a$ & snp$_b$ Critical region Null region Ideal

Null regions

GSS Quantification of Gain

Probability of selecting i controls from total…

…given proportion of
controls in null hypothesis region
(if they were the same distribution)

$$P_{\text{GSS}}(g_1, g_2) := \min_{(x_0,x_1)} \max_{(\pi_0,\pi_1)\in\mathcal{H}_0} \sum_{i=0}^{x_0} \binom{t_0}{i} \pi_0^i (1-\pi_0)^{t_0-i} \sum_{j=x_1}^{t_1} \binom{t_1}{j} \pi_1^j (1-\pi_1)^{t_1-j}, \tag{2}$$

Probability of selecting j cases…

…given proportion of
CASES in null hypothesis region

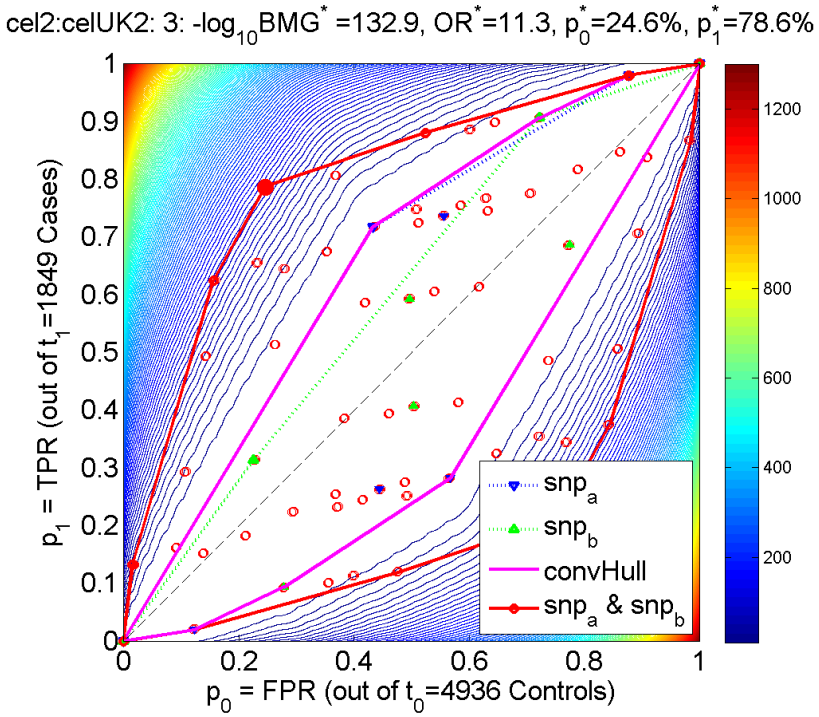
Celiac data

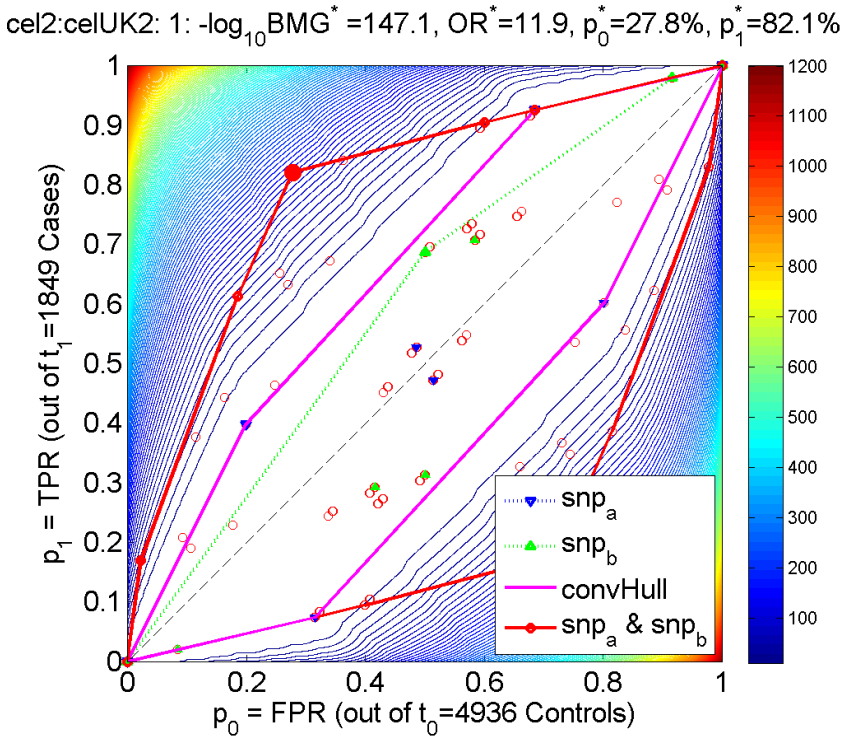
Courtesy of David van Heel, ICMS Queen Mary, University of London and M. Inouye, University of Melbourne

Source papers:

van Heel, D. A., et al., Nat. Genet., *39, 827–829, 2007*

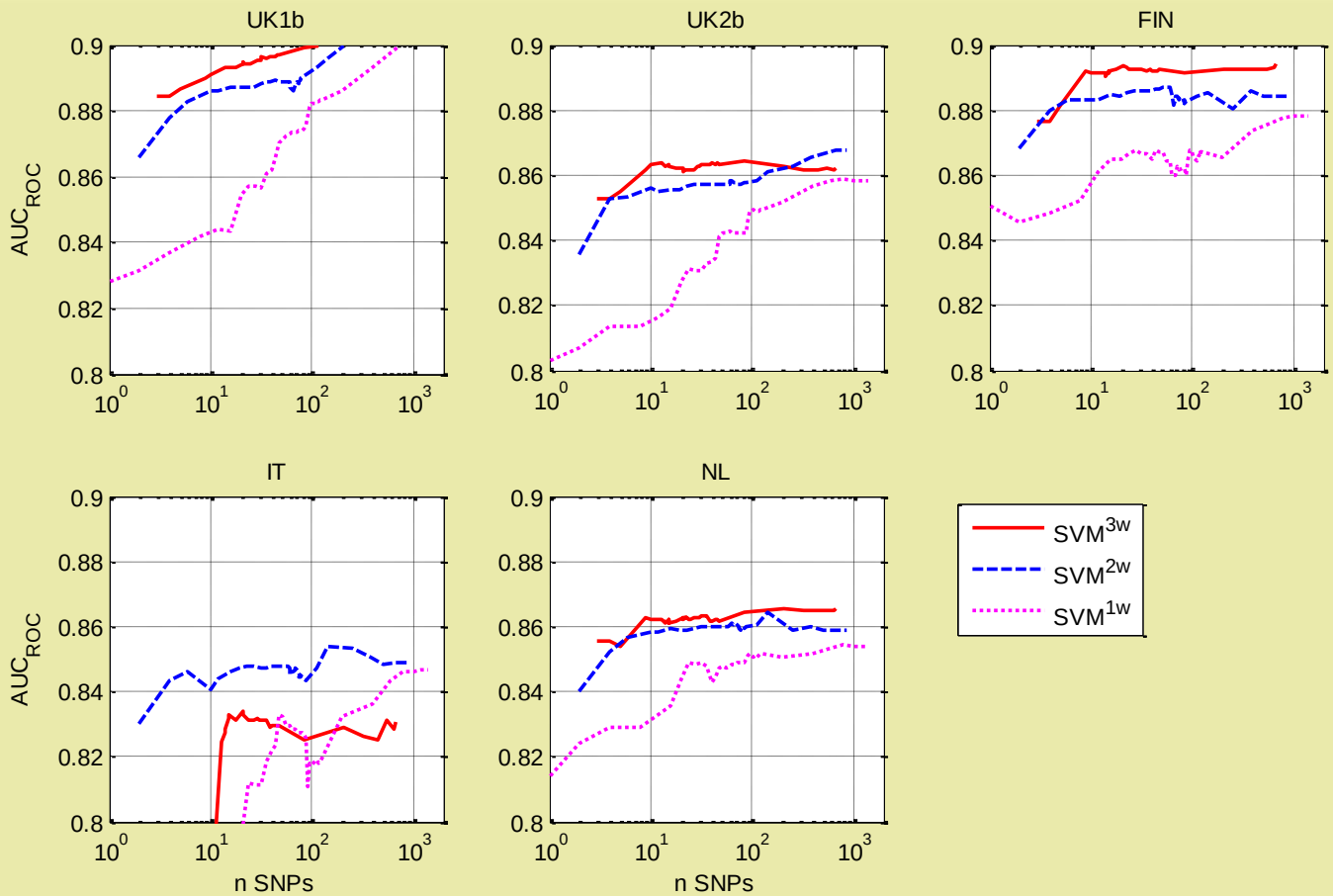
Dubois, P. C. A., et al., Nat. Genet., *42,295–304, 2010.*





Replication of risk prediction for Celiac Disease

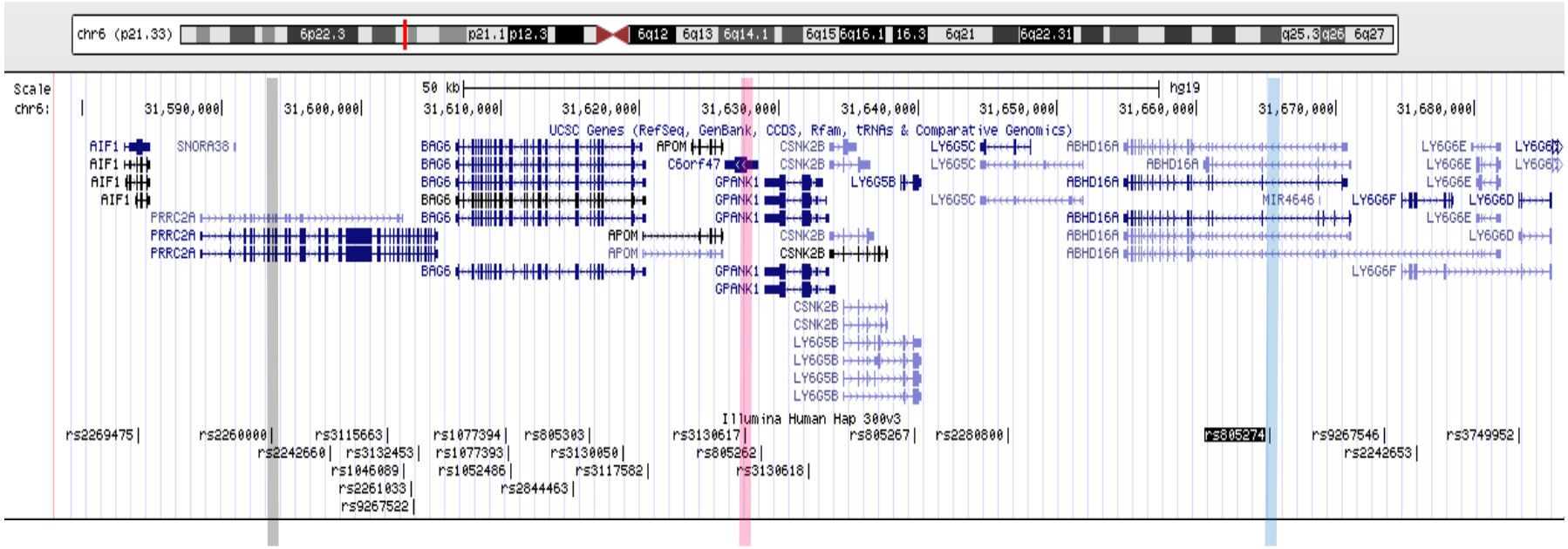
AUC for SVM predictors trained on sets of singletons, pairs or triplets of SNPs on UK1-discovery cohort, respectively



A novel hypothesis for Celiac Disease risk ready for follow-up verification

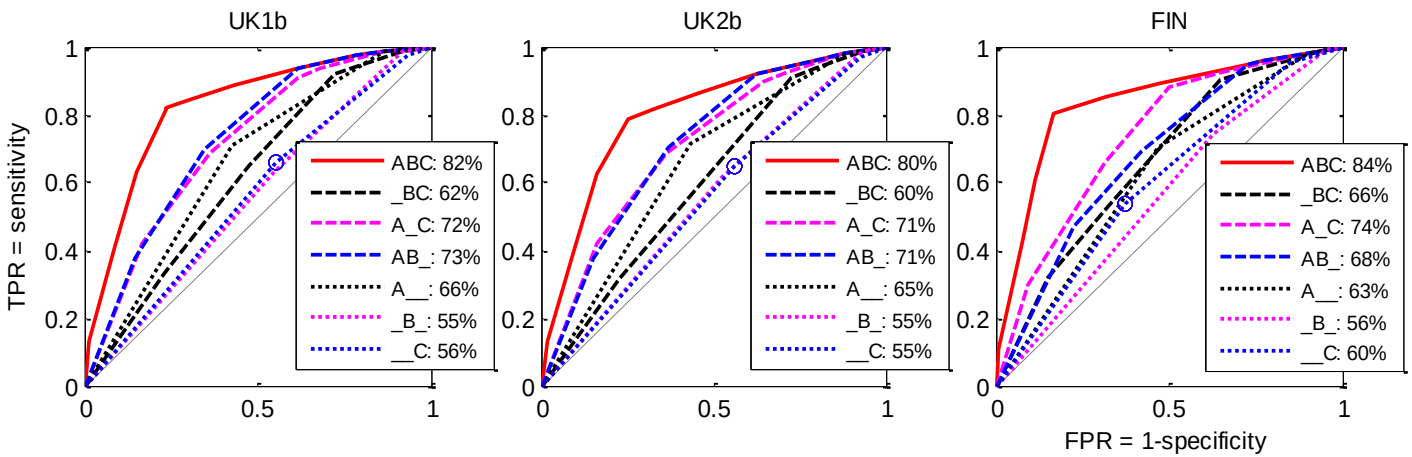
NICTA

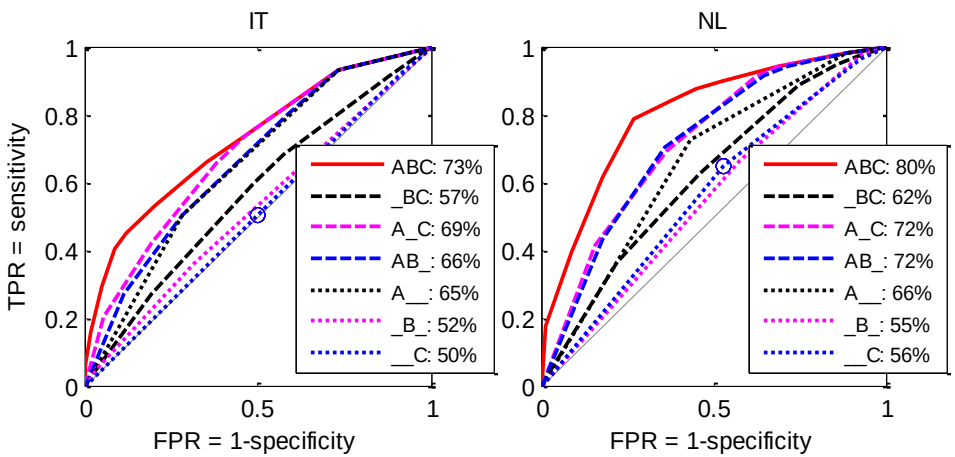
Triplet rs2260000, rs3130617, rs805274



A B C

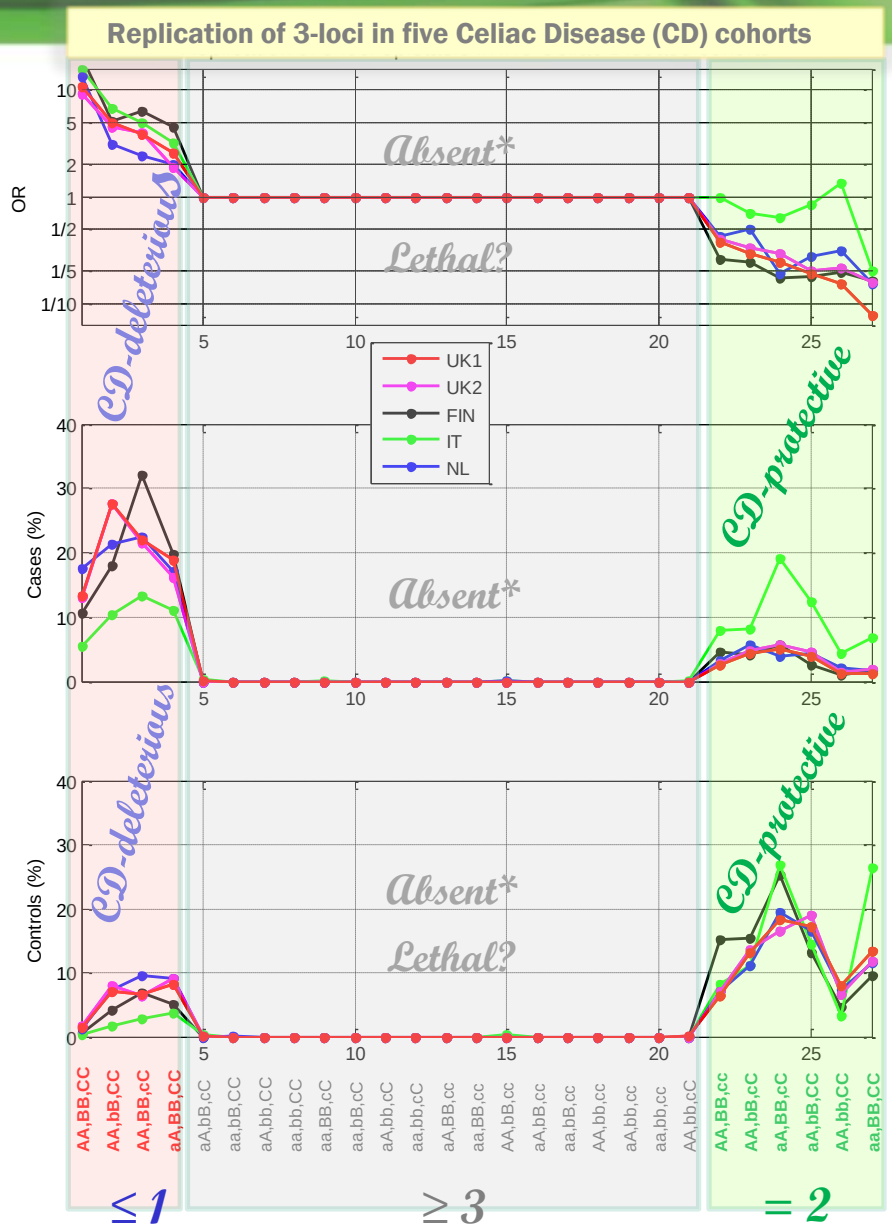
Replication of ROC curves for the triplet





		MAF UK1	MAF UK2	MAF FIN	MAF IT	MAF NL
A	rs2260000 chr6:31,593,476	0.35	0.34	0.31	0.49	0.34
B	rs3130617 chr6:31,627,523	0.27	0.27	0.21	0.17	0.25
C	rs805274 chr6:31,665,194	0.26	0.25	0.39	0.29	0.27

Replication of disease odds



* Such an "absence" of haplotypes we have verified in few more GWAS studies (of cancers)

A way forward

- **Biomedical validation**
  - demonstrate significant new biological insights
  - develop significant biomedical applications
- **Computational challenges**
  - improve timing and cost
  - integrated solutions (NGS)
  - unified pre-processing from raw data
- **Algorithmic**
  - epistasis on LD-block level
  - quantitative traits (20,000+ phenotypes)
  - predictions from complex genotypes (machine learning)

Conclusions

- Exhaustive search is necessary - no known efficient alternative.
- Exhaustive search of 2-way / 3-way interactions in GWAS is feasible & affordable, both computationally & algorithmically.
- Needs of more advanced hardware for 3-way GWIS analysis.
- GWIS provides a convenient extensible framework for quick prototyping of new statistical filters.

Acknowledgements

- Many thanks to G. Macintyre, K. Klotzbücher, H. Naeem, I. Lichtenstein, R.M. Campbell, A. Zarnegar, E. Kikianty, J. Markham, R. Mukhtar, G. Abraham, L. Stern and M. Inouye for help in this work.

- Thanks to D. A. van Heel and K. Hunt (Queen Mary University in London) for supplying the Celiac Disease data and help with analysis.

- Many thanks to Mark Berger and NVIDIA for providing Tesla K20 for experiments

- This work was supported by the Australian Research Council, and by the NICTA Victorian Research Laboratory. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications, and the Digital Economy, and the Australian Research Council through the ICT Centre of Excellence program.

- This research was undertaken with the assistance of resources provided at the Multi-modal Australian Sciences Imaging and Visualisation Environment (MASSIVE) through the National Computational Merit Allocation Scheme supported by the Australian Government.

Selected References

1. B. Goudey, D. Rawlinson, Q. Wang, F. Shi, H. Ferra, R. M. Campbell, L. Stern, M. T. Inouye, C. Soon Ong, A. Kowalczyk, GWIS - Model-free, Fast and Exhaustive Search for Epistatic Interactions in Case-Control GWAS, BMC Genomics 2013, accepted.
2. Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NLS, Yu W: BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. Am J Hum Genet 2010, 87(3):1-4.
3. van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, Inouye M, et al., A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21, Nat. Genet., *39, 827–829, 2007*
4. Brinza D, Schultz M, Tesler G, Bafna V: RAPID detection of gene-gene interactions in genome-wide association studies. Bioinformatics 2010, 26(22):2856-2862.
5. Dubois P. C. A., Trynka G., Franke L., A Hunt K.A., et al., Multiple common variants for celiac disease influencing immune gene expression, Nat. Genet., *42,295–304, 2010.*
6. Kam-Thong T, Putz B, Karbalai N, Muller-Myhsok B, Borgwardt K: Epistasis detection on quantitative phenotypes by exhaustive enumeration using GPUs. Bioinformatics 2011, 27(13):i214-i221.
7. Yung LS, Yang C, Wan X, Yu W: GBOOST: a GPU-based tool for detecting gene-gene interactions in genome-wide case control studies. Bioinformatics 2011, 27(9):1309-1310.
8. Hemani G, Theocharidis A, Wei W, Haley C: EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. Bioinformatics 2011, 27(11):1462-1465.
9. Hu X, Liu Q, Zhang Z, Li Z, Wang S, He L, Shi Y: SHEsisEpi, a GPU-enhanced genome-wide SNP-SNP interaction scanning algorithm, efficiently reveals the risk genetic epistasis in bipolar disorder. Cell Res 2010, 20(7):854-857.
10. The Wellcome Trust Case-Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007, 447(7145):661-678.
11. Li Chen, G. Yu, D.J.Miller,L. Song, C.Langefeld. D. Herrington, Y. Liu, Y Wang, A Ground Truth Based Comparative Study on Detecting Epistatic SNPs, Proc. IEEE Conf Bioinfotmatics Biomed, 2009 November 1, 26-31. Doi.10.1109/BIBMW.2009.5332132
12. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, Maller J, Sklar P, De Bakker P, Daly M, et al.: PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 2007, 81(3):559-575.
13. Prabhu S, Pe'er I: Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. Genome Res 2012, 22:2230-2240.
14. Q. Wang, F. Shi, A.C. Kowalczyk, D. Rawlinson, B. Goudey, R. M Campbell, H. Ferra and A. Kowalczyk (2014), GWIS_FI : a GPU Functional Interface Enables Exhaustive Search for Bivariate Interaction in Case-Control GWAS in Minutes, submitted.