and **Supplementary Table 4**). Many proteins were either overrepresented or underrepresented in each of the protease data sets, and clustering showed that enzyme specificity had the most influence on the results. Some examples within the top 1,000 proteins showed that for specific proteins, one protease outperformed all the others (**Fig. 1c** and **Supplementary Fig. 3**). Our data demonstrated that quantitation based on both spectral counting and peptide intensity was indeed biased when solely relying on a single protease, and this bias affected even the most abundant proteins, sometimes by more than a factor of 1,000. Amino acid analysis revealed that proteins overrepresented in a data set obtained by a particular protease contained relatively more cleavage-specific residues for that protease (**Supplementary Fig. 3**). Our data stresses that the best proteotypic peptides are not necessarily tryptic, a finding that may affect other quantitative assays such as selected reaction monitoring as well.

Raw and processed mass spectrometry identification data are available through thegpm.org at ftp://ftp.proteomecentral.org/public/0/ice.0.e.

Note: Supplementary information is available at http://www.nature.com/doifinder/10.1038/nmeth.2031/.

ACKNOWLEDGMENTS
This work was supported by the PRIME-XS project, grant agreement 262067, funded by the European Union 7th Framework Programme; the Netherlands Proteomics Centre of the Netherlands Genomics Initiative; the Centre for Biomedical Genetics; the Netherlands Organization for Scientific Research (NWO) VIDI grant (700.10.429) and the Netherlands Bioinformatics Centre.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Mao Peng[1,2], Nadia Taouatas[1,2], Salvatore Cappadona[1,2], Bas van Breukelen[1,2], Shabaz Mohammed[1,2], Arjen Scholten[1,2] & Albert J R Heck[1,2]

[1]Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands. [2]Netherlands Proteomics Centre, Utrecht, The Netherlands.
e-mail: a.scholten@uu.nl or a.j.r.heck@uu.nl

1. Huttlin, E.L. *et al. Cell* **143**, 1174–1189 (2010).
2. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E.M. *Nat. Biotechnol.* **25**, 117–124 (2007).
3. Malmström, J. *et al. Nature* **460**, 762–765 (2009).
4. Schwanhäusser, B. *et al. Nature* **473**, 337–342 (2011).
5. Mallick, P. *et al. Nat. Biotechnol.* **25**, 125–131 (2007).
6. Gauci, S. *et al. Anal. Chem.* **81**, 4493–4501 (2009).
7. de Godoy, L.M. *et al. Nature* **455**, 1251–1254 (2008).

Improved linear mixed models for genome-wide association studies

To the Editor: The use of linear mixed models (LMMs) in genome-wide association studies (GWAS) is now widely accepted[1] because LMMs have been shown to be capable of correcting for several forms of confounding due to genetic relatedness, such as population structure and familial relatedness[1], and because recent advances have made them computationally efficient[1,2]. LMMs tackle confounding by using a matrix of pairwise genetic similarities to model the relatedness among subjects. The consensus until now has been that all available single-nucleotide polymorphisms (SNPs) should be used to determine these similarities[1]. Here, however, we show theoretically and experimentally that carefully selecting a small number of SNPs systematically increases power (that is, it jointly reduces false positives and false negatives), improves calibration (lessens inflation or deflation of the test statistic) and reduces computational cost.

Our approach is motivated by two considerations. First, an LMM with no fixed effects using genetic similarities constructed from a set of SNPs is mathematically equivalent to a linear regression of the SNPs on the phenotype (with weights integrated over independent normal distributions having the same variance—in particular, the genetic variance)[3]. That is, an LMM using a given set of SNPs for genetic similarity is equivalent to (Bayesian) linear regression using those SNPs as covariates to correct for confounding. In theory, this equivalence holds only for certain forms of genetic similarity matrices, such as the realized relationship matrix[2,3]. In practice, however, the realized relationship matrix and other measures of similarity, such as identity by state[1], yield very similar measures of association (**Supplementary Note 1**), and thus our demonstration is quite general.

Second, regardless of the form of regression used for GWAS, the significance of SNP-phenotype association should be determined by conditioning on exactly those SNPs that are associated with the phenotype. These SNPs include causal SNPs, or those nearby that tag causal SNPs, and SNPs that are associated by way of confounding (for example, because of population structure). By conditioning on causal or tagging SNPs, we reduce the noise in the assessment of the association[4]. By conditioning on SNPs associated because of confounding, we control for such confounding[5]. Moreover, if a SNP is unrelated to the phenotype, it should not be in the conditioning set. In the particular case in which we use Bayesian linear regression for GWAS, the inclusion of unrelated SNPs in the genetic similarity matrix decreases the relative influence of each SNP on the phenotype (because all SNP weights share the same prior distribution whose variance—the genetic variance in the LMM view—is estimated from the data). The decrease in influence leads to incomplete correction for confounding and hence inflated test statistics and reduced power. We refer to this phenomenon as 'dilution.'

To identify SNPs that satisfy these principles, we developed a simple heuristic that yields improved power and calibration. First, we order SNPs by their linear-regression P values from lowest to highest. Then we construct genetic similarity matrices with an increasing number of SNPs as previously ordered until we find the first minimum in λ_{GC} (the genomic control factor). In practice, the number of SNPs selected is typically smaller than the number of individuals analyzed, a condition that can be exploited by an existing algorithm, FaST-LMM, to yield large computational savings[2].

The equivalence between the LMM and Bayesian linear regression also implies that, when a given SNP is being tested, that SNP should be excluded from the computation of genetic similarity to avoid using it as a covariate. Including the SNP would make the log likelihood of the null model higher than it should be and lead to deflation of the test statistic and loss of power. We call this phenomenon 'proximal contamination'. In addition to the SNP being tested, we also exclude those SNPs in close proximity (for example, within 2 centimorgans), as linkage disequilibrium will lead to a similar deflation and loss of power. A naive algorithm for excluding these from the similarity matrix is computationally expensive, so we developed a speedup (**Supplementary Note 2**). Together, the linear-regression scan to select SNPs for inclusion in the matrix

Table 1 | Comparison of calibration, power and computational costs on a GWAS of Crohn's disease

Algorithm parameters				Algorithm performance					
Algorithm	**SNP selection method**	**No. SNPs for matrix**	**Proximal contamination avoided?**	λ_{GC}	**No. false positives**	**No. false negatives**	**Runtime without speedup (min)**	**Runtime with speedup (min)**	**Memory usage (GB)**
FaST-LMM-Select	Select	310	Yes	1.08	0	1	1.3×10^3	45	<1
FaST-LMM (all)	All	All	Yes	1.09	2	2	4.0×10^6	4,567	86
FaST-LMM (orig 310)	Equally spaced	310	Yes	1.26	9	1	1.1×10^3	6	<1
FaST-LMM (orig 4,000)	Equally spaced	4,000	Yes	1.17	5	1	2.1×10^5	30	2
Traditional	All	All	No	0.97	2	6	4.2×10^1	NA	45

The original version of FaST-LMM, which used equally spaced SNPs to estimate genetic similarity, was evaluated using 310 SNPs (the same number used by FaST-LMM-Select) and 4,000 SNPs (as used in the original version of FaST-LMM (ref.2)). The five algorithms yielded substantially different *P* values (**Supplementary Fig. 1**), which in turn led to different SNPs being deemed significant (using the *P* value threshold of 5×10^{-7} (ref. 6)). Previous studies were used to determine the gold standard in order to label the false positive and false negative loci (**Supplementary Table 1**). Details of the analysis are described in the **Supplementary Methods**.

along with the efficient removal of the test SNPs and those nearby constitute our new approach, FaST-LMM-Select.

When applied to Wellcome Trust data for Crohn's disease[6] (**Table 1, Supplementary Fig. 1, Supplementary Table 1** and **Supplementary Methods**) that includes family members and non-Caucasians, FaST-LMM-Select yielded slightly less inflation, fewer false positives and fewer false negatives (due to lack of dilution) compared to the use of all SNPs while accounting for proximal contamination. When all SNPs were used, proximal contamination had a dramatic effect on calibration and false positives even though correction for it excluded (on average) only 516 of the available 356,441 SNPs from the genetic similarity matrix. Compared with the original version of FaST-LMM, wherein equally spaced SNPs were used to reduce computational demands, FaST-LMM-Select had far better calibration and fewer false positives. FaST-LMM-Select also performed well on synthetic data (**Supplementary Note 1**) and other real cohorts with substantial genetic structure (**Supplementary Note 3**).

FaST-LMM-Select is available at http://mscompbio.codeplex.com/.

Note: Supplementary information is available at http://www.nature.com/doifinder/10.1038/nmeth.2037.

ACKNOWLEDGMENTS
We thank J. Carlson for help with tools to manage and analyze the data and P. Palamara for cataloging the positions and genetic distances of SNPs in the data for Crohn's disease. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk/. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475. E.E. is supported by US National Science Foundation grants 0916676 and 1065276 and by US National Institutes of Health grants K25-HL080079, U01-DA024417, P01-HL30568 and P01-HL28481.

AUTHOR CONTRIBUTIONS
J.L., C.L. and D.H. designed and performed the research, contributed analytic tools, analyzed data and wrote the paper. C.M.K. and R.I.D. contributed analytic tools. E.E. helped to write the paper.

COMPETING FINANCIAL INTERESTS
The authors declare competing financial interests: details are available at http://www.nature.com/doifinder/10.1038/nmeth.2037.

Jennifer Listgarten[1,5], Christoph Lippert[1,2,5], Carl M Kadie[3], Robert I Davidson[3], Eleazar Eskin[4] & David Heckerman[1,5]

[1]Microsoft Research, Los Angeles, California, USA. [2]Max Planck Institutes Tübingen, Tübingen, Germany. [3]Microsoft Research, Redmond, Washington, USA. [4]University of California Los Angeles, Los Angeles, California, USA. [5]These authors contributed equally to this work.
e-mail: jennl@microsoft.com, christoph.lippert@tuebingen.mpg.de or heckerma@microsoft.com

1. Kang, H.M. *et al. Nat. Genet.* **42**, 348–354 (2010).
2. Lippert, C. *et al. Nat. Methods* **8**, 833–835 (2011).
3. Hayes, B.J., Visscher, P.M. & Goddard, M.E. *Genet. Res. (Camb.)* **91**, 47–60 (2009).
4. Hoggart, C.J., Whittaker, J.C., Iorio, M.D. & Balding, D.J. *PLoS Genet.* **4**, e1000130 (2008).
5. Setakis, E., Stirnadel, H. & Balding, D.J. *Genome Res.* **16**, 290–296 (2006).
6. Wellcome Trust Case Control Consortium. *Nature* **447**, 661–678 (2007).