Deletion polymorphism upstream of *IRGM* associated with altered *IRGM* expression and Crohn's disease

Steven A McCarroll, Alan Huett , Petric Kuballa, Shannon D Chilewski,
Aimee Landry, Philippe Goyette, Michael C Zody, Jennifer L Hall,
Steven R Brant, Judy H Cho, Richard H Duerr, Mark S Silverberg, Kent D Taylor,
John D Rioux, David Altshuler, Mark J Daly, and Ramnik J Xavier

Supplementary Information:

Supplementary Methods

Robust breakpoint assay for typing 20-kb deletion polymorphism upstream of *IRGM*

Introduction

We described in this work a 20-kb deletion polymorphism at *IRGM*, in strong LD with SNPs at *IRGM* previously shown to be associated with Crohn's disease. Typing this variant in clinical cohorts will be an important need in efforts to evaluate the association of this polymorphism with other clinical phenotypes, and in efforts to definitively identify the causal variant at the locus. We therefore went to extensive effort to develop a robust typing assay that yielded genotypes approaching 100% completeness and accuracy.

Genetic fine-mapping (the effort to identify a potential causal variant among several polymorphisms that are in linkage disequilibrium (LD) with one another) is particularly dependent on assays that are close to 100% complete and accurate. Even a small amount of missing data can lead to the incorrect inference that one variant appears to be "more associated" or "less associated" than others with which it is in fact in perfect LD. More important, in conditional association analysis (in which one analyzes the phenotypic association of variant A, conditional on the genotype at linked variant B), only a small fraction of the individuals in a study cohort may carry the rare informative haplotypes on which measurement of a conditional association signal is based; in such studies, even a low rate of genotyping error can obscure a real signal or create a false signal, particularly when genotyping error is confounded by case/control status (as when DNA samples from affected and unaffected individuals has different clinical origins or are genotyped on different plates, a common phenomenon known as "differential bias" and described in Clayton, D.G., *et al., Nat Genet* 2005).

We developed a breakpoint assay based on the specific sequence of the molecular lesion at this locus (**Fig. 1c**). In this assay, distinct amplicons are generated from the reference and deletion alleles (in the same reaction well) and separately detected by means of TaqMan probes with distinct fluorophores. The amplicon specific to the deletion allele is generated from primer sequences that flank the deletion breakpoint; the amplicon specific to the reference allele is generated from within the 20-kb deleted segment. Generation of both amplicons in the same reaction well is important: in the absence of such an internal control, pipetting error, detection error, and air bubbles can lead to genotype error. Development of such a multiplex breakpoint assay proved challenging because (as is the case for many structural variants), one end of this deletion resides in GC-poor sequence within a high-copy repeat.

Robustness of the assay requires that both amplicons be generated with similar kinetics, such that one amplicon does not consume too large a fraction of the shared pool of free nucleotides before the other amplicon is sufficiently amplified for robust detection. Assay development therefore required extensive optimization of primer sequences, annealing temperature, and primer and probe concentrations. (While the optimization process is not described here, we would be glad to correspond with scientists trying to develop analogous typing assays at other structurally variant loci.)

Validation of assay results

We used two approaches to validate the results of this assay:

1. *Concordance with an independent method.* We used the assay to genotype the 270 HapMap individuals, and compared the resulting genotypes to genotypes obtained in the same individuals by an independent method (Affymetrix SNP 6.0 array, described in McCarroll *et al.*, *Nature Genetics,* in press). The breakpoint assay described here yielded 100% complete data (270 genotypes) and 100% concordance with genotypes from the independent, array-based approach.

2. *Unequivocal LD relationship with a tightly linked variant typed by an independent method.* We used this assay to type the 20-kb deletion polymorphism in a Crohn's cohort of 685 individuals. We obtained 99.4% complete data (100% complete after uncalled samples were analyzed a second time) and perfect correlation ($r^2 = 1.0$) with genotypes for a tightly linked SNP (rs13361189) that was genotyped using an independent method (Sequenom).

Materials – reagents:

1. TaqMan Master Mix (Applied Biosystems)
2. Optical reaction plates for real-time PCR (Applied Biosystems)
3. Oligonucleotide primers and probes (from Integrated DNA Technologies), with the following sequences:

Name	**Sequence**
IRGM_insAllele_f0	TATCTGCTTGGCTGGGAAAC
IRGM_insAllele_r0	CCATACAGCCCTTTGCCTTA
IRGM_insAllele_FAM	FAM – TCCCTTGCTGAGGAGACCCACTC – BHQ
IRGM_delAllele_f8	TTAAAATTTAGGAATAAATTTAATAATTTAATAATAAAATAG
IRGM_delAllele_r3	GTGGGAGGTCAGCAGAAAAC
IRGM_delAllele_HEX	HEX - TTCCAGGCTTTTTCTGTGACCATGTC - BHQ

Materials -- equipment:

Applied Biosystems 7900HT instrument

Procedure:

1. Create an experimental reaction plate with 2 ul of each DNA sample (at a concentration of 10 ng/ul).

2. Create a master reaction mix containing the following volumes of each reagent. (Volumes shown are per 100 samples typed, and should be scaled appropriately to reflect the actual number of samples typed.)

Reagent	**Volume (ul)**
TaqMan master mix	1,000
Distilled H20	729
Primer IRGM_insAllele_f0 (50 uM)	2
Primer IRGM_insAllele_r0 (50 uM)	2
Probe IRGM_insAllele_FAM (50 uM)	2
Primer IRGM_delAllele_f8 (50 uM)	50
Primer IRGM_delAllele_r3 (50 uM)	10
Probe IRGM_delAllele_HEX (50 uM)	5
Total volume	1,800

Note that the concentration of primers and probe for the deletion-allele amplicon are much greater than those for the reference-allele amplicon. This helps to give the two amplicons similar amplification/detection kinetics, compensating for the GC-poor quality of the sequence around the left deletion breakpoint.

3. Add 18 ul of master reaction mix to each well. Make sure that no air bubble separates the added reaction mix from the DNA sample.

4. Centrifuge plate (2 min at 2000 rpm) to remove air bubbles.

5. Perform real-time PCR, detecting FAM and HEX fluorescence signals, using the following cycling protocol:

	50C	2:00
	95C	10:00
x 40 cycles	95C	0:15
	55C	1:30

Critical steps

1. Visual examination and manual curation of the real-time amplification curves is always recommended, as the analysis software that accompanies real-time PCR instruments occasionally makes false-positive determinations that a threshold cycle has been reached, when visual examination of the amplification curve (and comparison to the curves for other samples on the same plate) makes clear that no significant amplification has taken place.

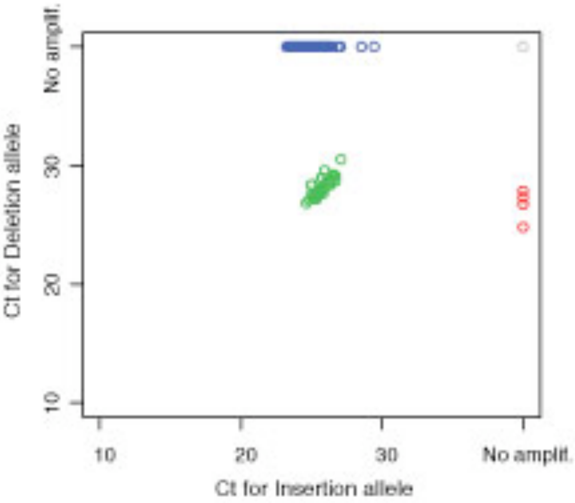
2. It is always recommended that, when running this assay (or any other typing assay) in a new lab or on a new instrument, that samples of known genotype be included on the

reaction plate. The final section of this supplementary method ("Reference genotypes") contains reference genotypes obtained for this polymorphism in the HapMap samples (available from Coriel Laboratories).

3. Although in our experience the assay routinely produces highly complete and accurate data, it is nonetheless recommended that in analyses involving fine-mapping, the assay be performed twice on each sample both: (1) to fill-in any missing or equivocal genotypes, and (2) that the concordance of the two experimental replicates be verified, and subsequent analysis limited to genotypes that were identical between replicates (in our experience, 99.5-100% of all genotypes obtained).

Anticipated results

The following plot shows an example of the results of this assay, for a 384-well reaction plate including one empty well. Each DNA sample is represented by a point. On the horizontal axis is the threshold amplification cycle for the amplicon arising from the rererence allele. On the vertical axis is the threshold amplification cycle for the amplicon arising from the deletion allele. The three genotype classes correspond to the genotypes of +/+ (blue), +/- (green), and -/- (red). Outlier samples (grey) should be considered to have an indeterminate genotype.



Reference genotypes

Following are reference genotypes generated using this assay, for the 270 HapMap samples (available from Coriel Laboratories); 2 indicates a genotype of +/+; 1 indicates a genotype of +/-; 0 indicates a genotype of -/-.

Sample	Genotype
NA06985	2
NA06991	2
NA06993	2
NA06994	2
NA07000	2
NA07019	2
NA07022	2
NA07029	2
NA07034	2
NA07048	2
NA07055	2
NA07056	2
NA07345	2
NA07348	2
NA07357	2
NA10830	2
NA10831	2
NA10835	2
NA10838	2
NA10839	2
NA10846	2
NA10847	2
NA10851	2
NA10854	2
NA10855	2
NA10856	2
NA10857	2
NA10859	2
NA10860	2
NA10861	2
NA10863	2
NA11829	2
NA11830	2
NA11831	2
NA11832	2
NA11839	2
NA11840	2
NA11881	2
NA11882	2
NA11992	2
NA11993	2
NA11994	2
NA11995	1
NA12003	2
NA12004	2
NA12005	2
NA12006	2

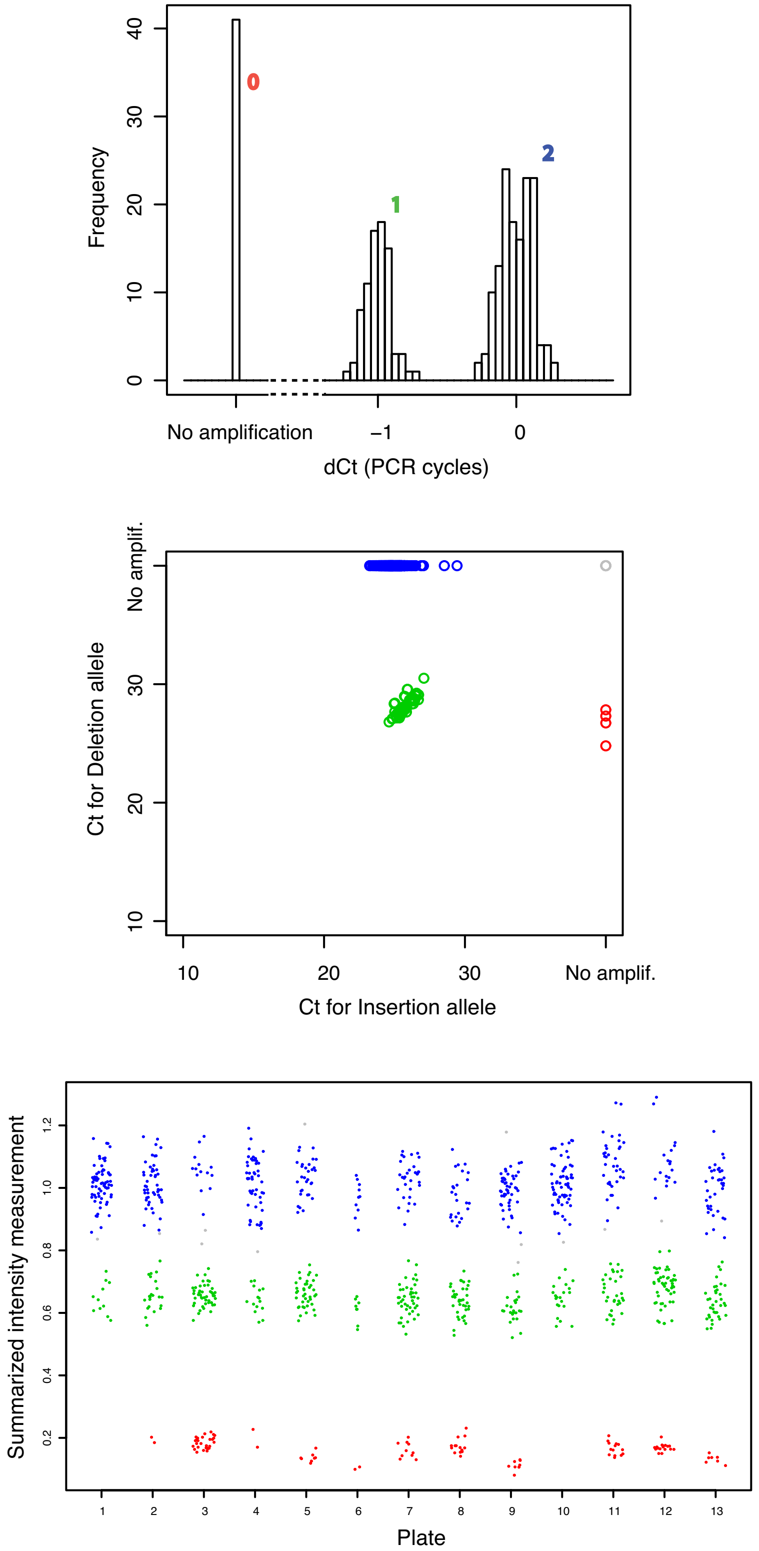
```
NA12043 2
NA12044 2
NA12056 2
NA12057 2
NA12144 2
NA12145 2
NA12146 2
NA12154 2
NA12155 2
NA12156 2
NA12234 2
NA12236 1
NA12239 2
NA12248 2
NA12249 2
NA12264 2
NA12707 2
NA12716 2
NA12717 2
NA12740 2
NA12750 2
NA12751 2
NA12752 2
NA12753 2
NA12760 2
NA12761 2
NA12762 2
NA12763 2
NA12801 2
NA12802 2
NA12812 2
NA12813 2
NA12814 2
NA12815 2
NA12864 2
NA12865 2
NA12872 2
NA12873 2
NA12874 1
NA12875 2
NA12878 2
NA12891 2
NA12892 2
NA18500 0
NA18501 1
NA18502 0
NA18503 1
NA18504 1
```

```
NA18505 1
NA18506 2
NA18507 2
NA18508 2
NA18515 1
NA18516 0
NA18517 1
NA18521 2
NA18522 1
NA18523 1
NA18852 2
NA18853 1
NA18854 1
NA18855 1
NA18856 0
NA18857 0
NA18858 1
NA18859 2
NA18860 1
NA18861 1
NA18862 0
NA18863 1
NA18870 2
NA18871 1
NA18872 2
NA18912 1
NA18913 2
NA18914 2
NA19092 1
NA19093 0
NA19094 1
NA19098 0
NA19099 1
NA19100 1
NA19101 2
NA19102 1
NA19103 2
NA19116 1
NA19119 0
NA19120 1
NA19127 0
NA19128 2
NA19129 1
NA19130 0
NA19131 1
NA19132 1
NA19137 0
NA19138 2
```

NA19139	1
NA19140	0
NA19141	1
NA19142	0
NA19143	2
NA19144	1
NA19145	2
NA19152	1
NA19153	2
NA19154	2
NA19159	0
NA19160	1
NA19161	0
NA19171	0
NA19172	0
NA19173	0
NA19192	0
NA19193	1
NA19194	1
NA19200	0
NA19201	1
NA19202	0
NA19203	1
NA19204	2
NA19205	1
NA19206	0
NA19207	1
NA19208	1
NA19209	0
NA19210	1
NA19211	1
NA19221	2
NA19222	1
NA19223	2
NA19238	1
NA19239	2
NA19240	1
NA18524	2
NA18526	2
NA18529	2
NA18532	2
NA18537	0
NA18540	1
NA18542	1
NA18545	0
NA18547	0
NA18550	1
NA18552	1

NA18555	0
NA18558	2
NA18561	1
NA18562	1
NA18563	1
NA18564	1
NA18566	1
NA18570	0
NA18571	1
NA18572	0
NA18573	1
NA18576	0
NA18577	2
NA18579	2
NA18582	1
NA18592	2
NA18593	1
NA18594	1
NA18603	2
NA18605	0
NA18608	2
NA18609	1
NA18611	1
NA18612	1
NA18620	2
NA18621	2
NA18622	1
NA18623	1
NA18624	1
NA18632	1
NA18633	2
NA18635	1
NA18636	2
NA18637	1
NA18940	1
NA18942	1
NA18943	0
NA18944	2
NA18945	0
NA18947	0
NA18948	2
NA18949	0
NA18951	2
NA18952	2
NA18953	0
NA18956	1
NA18959	1
NA18960	2

NA18961	2
NA18964	1
NA18965	2
NA18966	2
NA18967	1
NA18968	2
NA18969	2
NA18970	1
NA18971	2
NA18972	2
NA18973	1
NA18974	2
NA18975	2
NA18976	1
NA18978	1
NA18980	0
NA18981	2
NA18987	1
NA18990	0
NA18991	2
NA18992	2
NA18994	1
NA18995	0
NA18997	1
NA18998	1
NA18999	1
NA19000	0
NA19003	2
NA19005	1
NA19007	2
NA19012	1



a. Interrogation of copy-number variable segment by quantitative PCR
(two-color TaqMan assay, comparing amplification of this segment to that of a control, two-copy locus)

Data shown: 270 HapMap samples, three experimental replicates averaged by median polish

b. Breakpoint assay
(two-color TaqMan assay, detecting an amplicon specific to each structural allele)

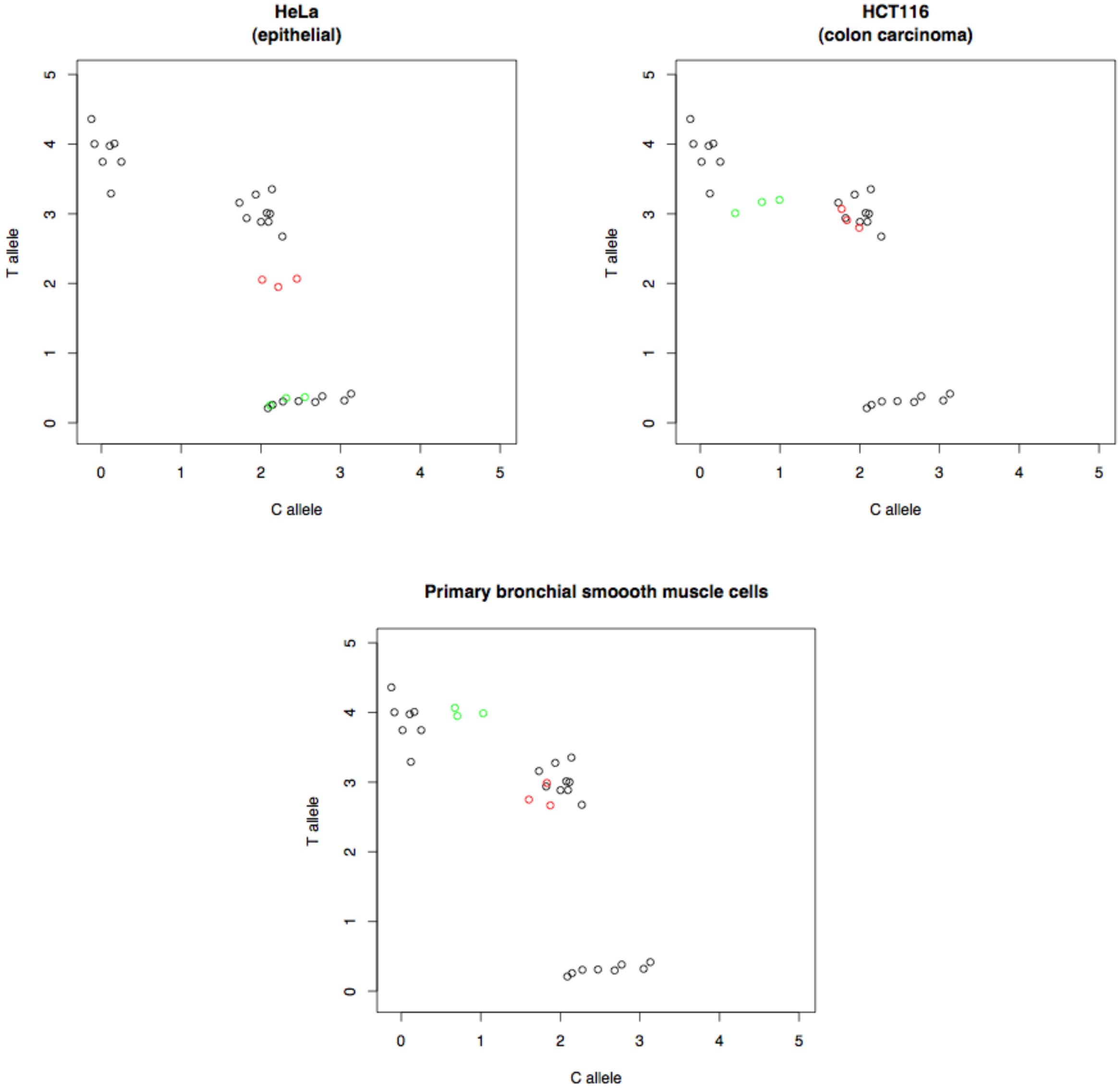
Data shown: 384 IBD cases and controls.

c. SNP6.0 array
(summarized intensity measurements from six copy-number probes across deleted segment)

Data shown: 13 plates of 20-96 samples each.

Supplementary Figure 1

Methods of interrogating the 20 kb *IRGM* deletion polymorphism in this study, with representative examples of the data from each assay.



Supplementary Figure 2. Results of TaqMan SNP assay for rs10065172 in genomic DNA (red circles) and cDNA (green circles) from HeLa S3, HCT116, and primary smooth muscle cells. Black circles show results for a control set of genomic DNA samples (drawn from HapMap YRI population sample).

Supplementary Figure 3
IRGM-C-FLAG expression construct

IRGM(a) [UniprotKB entry: A1A4Y4]

cDNA (546 bp):

```
atggaagccatgaatgttgagaaagcctcagcagatgggaacttgccagaggtgatctctaacatcaagg
agactctgaagatagtgtccaggacaccagttaacatcactatggcaggggactctggcaatgggatgtc
caccttcatcagtgcccttcgaaacacaggacatgagggtaaggcctcacctcctactgagctggtaaaa
gctacccaaagatgtgcctcctatttctcttcccacttttcaaatgtggtgttgtgggacctgcctggca
cagggtctgccaccacaaccctggagaactacctgatggaaatgcagttcaaccggtatgacttcatcat
ggttgcatctgcacaattcagcatgaatcatgtgatgcttgccaaaaccgctgaggacatgggaaagaag
ttctacattgtctggaccaagctagacatggacctcagcacaggtgccctcccagaagtgcagctactgc
agatcagagaaaatgtcctggaaaatctccagaaggagcgggtatgtgaatactaa
```

Protein (181 aa):

MEAMNVEKASADGNLPEVISNIKETLKIVSRTPVNITMAGDSGNGMSTFISALRNTGHEGKASPP
TELVKATQRCASYFSSHFSNVVLWDLPGTGSATTTLENYLMEMQFNRYDFIMVASAQFSMNHV
MLAKTAEDMGKKFYIVWTKLDMDLSTGALPEVQLLQIRENVLENLQKERVCEY

IRGM(a)-3x-FLAG (tag at C-terminus):

MEAMNVEKASADGNLPEVISNIKETLKIVSRTPVNITMAGDSGNGMSTFISALRNTGHEGKASP
PTELVKATQRCASYFSSHFSNVVLWDLPGTGSATTTLENYLMEMQFNRYDFIMVASAQFSMNH
VMLAKTAEDMGKKFYIVWTKLDMDLSTGALPEVQLLQIRENVLENLQKERVCEY-LRPL-
DYKDDDDK-G-DYKDDDDK-G-DYKDDDDK*
[**bold = IRGM(a)**, underlined = FLAG-tag]

Supplementary Table 1
Primer, probe, and siRNA sequences used

Quantitative PCR assay for genotyping *IRGM* deletion (a two-color TaqMan assay, in which the affected locus and a control, two-copy locus were simultaneously amplified and detected using TaqMan probes)

Affected locus	
Forward primer	TATCTGCTTGGCTGGGAAAC
Reverse primer	CCATACAGCCCTTTGCCTTA
Probe	*FAM*–TCCCTTGCTGAGGAGACCCACTC–*BHQ*
Control locus	
Forward primer	CCCTTCTCAGCGGTGTCATC
Reverse primer	ACAGACCGTCTGGGCGC
Probe	*VIC*–TTCGCGTTTCCGCAAGAT–*MGBFQ*

Breakpoint assay for genotyping *IRGM* deletion (a two-color Taqman assay utilizing simultaneous amplification of amplicons specific to the insertion and deletion alleles)

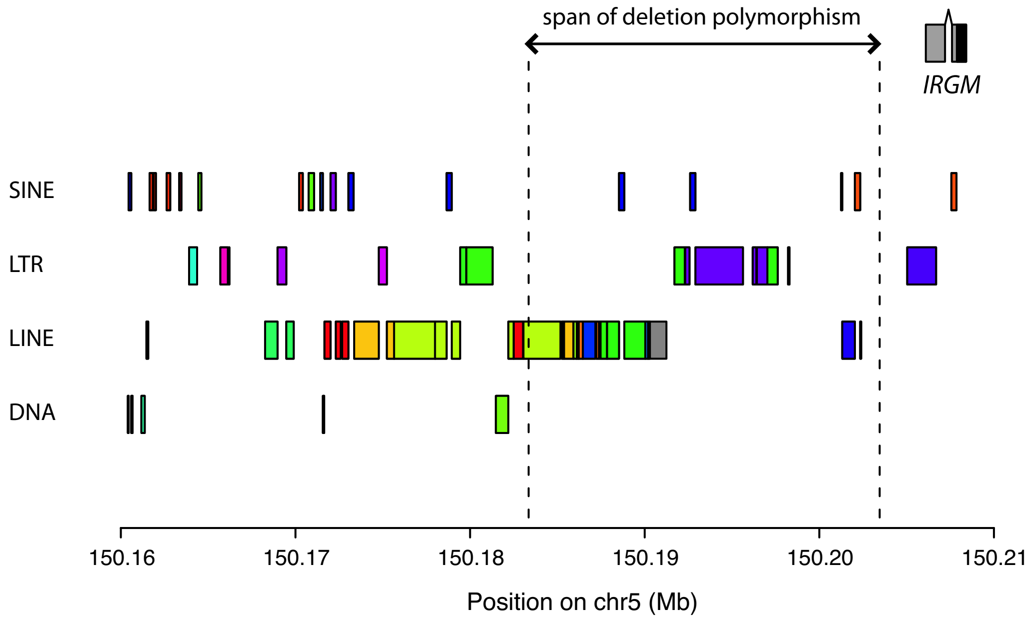
Insertion allele	
Forward primer	TATCTGCTTGGCTGGGAAAC
Reverse primer	CCATACAGCCCTTTGCCTTA
Probe	*FAM*–TCCCTTGCTGAGGAGACCCACTC–*BHQ*
Deletion allele	
Forward primer	TTAAAATTTAGGAATAAATTTAATAATTTAATAATAAAATAG
Reverse primer	GTGGGAGGTCAGCAGAAAC
Probe	*HEX*–TTCCAGGCTTTTTCTGTGACCATGTC–*BHQ*

SNP assay (Sequenom platform) for rs10065172

forward primer	ACGTTGGATGAGTCATACCGGTTGAACTGC
reverse primer	ACGTTGGATGCAAATGTGGTGTTGTGGGAC
probe	TGAACTGCATTTCCATCA

Sequences of siRNA duplexes:

si1 5’-UUUCGAAGGGCACUGAUGAAGGUGG-3’
si2 5’-AUUUCCAUCAGGUAGUUCUCCAGGG-3’
si3 5’-UGUCCUGGACACUAUCUUCAGAGUC-3’



Supplementary Note
Gene immigration and structural evolution at the *IRGM* locus

The genomic region upstream of *IRGM* shows extensive recent modification by retroposons as well as the deletion polymorphism described here. In the above figure, colored rectangles indicate sites of retroposon insertions, with colors representing the 33 types of retroposon that have invaded the genomic region upstream of *IRGM*, mostly along the primate lineage; retroposon sequences are shown on four lines reflecting their superfamily organization. Grey and black rectangles at top right indicate the genomic location of the core *IRGM* transcript (grey indicates 5’ UTR as described in ref[1]; black indicates protein-coding sequence).

IRGM appears to have arrived at its primate genomic location by either an extremely small translocation event or possibly a retroposition, as evidenced the absence of *IRGM*-like genes in the syntenic regions of mammalian genomes (outside the primate-rodent branch) that have contiguous conserved synteny across the locus but lack *IRGM*-like genes at the locus (dog[2] and horse [K. Lindblad-Toh, personal communication]). The region of the human genome containing *IRGM* is in a segment of conserved synteny with dog chromosome 4 and horse chromosome 14, which contain no evidence of *IRGM*. In mouse, the region containing human *IRGM* lies across a syntenic break caused by a chromosomal fission in the rodent lineage. The only *IRGM*-like genes in the opossum genome[3] lie on a different chromosome from the region syntenic to human *IRGM*, suggesting that the dog and horse genomes reflect the ancestral state of the locus.

Mouse *Irgm* lies within a distinct region whose human synteny is to the very distal tip of human 5q, but the gene itself and about 15 kb of upstream sequence are completely missing from the human genome at that location (with the exception of a few tRNA hits which appear to be conserved). The dog genome contains three loci similar to mouse *Irgm*, all within a 400 kb region at the far proximal end of dog chromosome 11; this region is in a conserved syntenic segment with respect to the mouse *Irgm* locus on chromosome 11 and human 5qter, which does not contain *IRGM*. Collectively, these data suggest that human *IRGM* resides in a different genomic context than its mammalian orthologs.

The primate copy of *IRGM* was present at its current genomic location at least before the divergence of new and old world monkeys, ~35 Mya [4], as it is present in marmoset (R. Gibbs, personal communication). The segment of mouse chromosome 18 adjacent to this locus contains a series of genes which appear homologous to the gene adjacent to mouse *Irgm* (*Tgtp*), suggesting that this event may have moved both genes here prior to the divergence of human and mouse, ~75 Mya[5]. However, the exact extent of the duplication is not knowable, as there is no sequence similarity between any sequence at the human *IRGM* locus (outside of the principal coding exon of *IRGM* itself) with any other locus in human or mouse.

Although human *IRGM* has a splice site in its 5' UTR and can also yield transcripts with alternative 3' exons [1], conservation of human *IRGM* with other *Irg* family members (including *Irgm*) is limited to the core coding exon of *IRGM*, consistent with a model in which the core coding exon of *IRGM* arrived at its primate genomic location in a primate-rodent ancestor prior to the subsequent evolution of its upstream sequence.

[1] Bekpen, C. *et al*. The interferon-inducible p47 (IRG) GTPases in vertebrates: loss of the cell autonomous resistance mechanism in the human lineage. *Genome Biol* **6**, R92 (2005).

[2] Lindblad-Toh, K. *et al*. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438, 803-819 (2005).

[3] Mikkelsen, T.S., *et al*. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequence. Nature 447, 167-77.

[4] Schrago, C.G. Time scale of New World primate diversification. *J Phys Anthropol* 132, 344-54.

[5] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520-562 (2002).