Genome Medicine

RESEARCH Open Access

# Complex host genetics influence the microbiome in inflammatory bowel disease

Dan Knights[1,2,3,4*], Mark S Silverberg[5†], Rinse K Weersma[6†], Dirk Gevers[2], Gerard Dijkstra[6], Hailiang Huang[7], Andrea D Tyler[5], Suzanne van Sommeren[6,8], Floris Imhann[6,8], Joanne M Stempak[5], Hu Huang[9], Pajau Vangay[9], Gabriel A Al-Ghalith[9], Caitlin Russell[3,10], Jenny Sauk[10], Jo Knight[11], Mark J Daly[2,12,13], Curtis Huttenhower[2,14] and Ramnik J Xavier[2,3,10*]

## Abstract

**Background:** Human genetics and host-associated microbial communities have been associated independently with a wide range of chronic diseases. One of the strongest associations in each case is inflammatory bowel disease (IBD), but disease risk cannot be explained fully by either factor individually. Recent findings point to interactions between host genetics and microbial exposures as important contributors to disease risk in IBD. These include evidence of the partial heritability of the gut microbiota and the conferral of gut mucosal inflammation by microbiome transplant even when the dysbiosis was initially genetically derived. Although there have been several tests for association of individual genetic loci with bacterial taxa, there has been no direct comparison of complex genome-microbiome associations in large cohorts of patients with an immunity-related disease.

**Methods:** We obtained 16S ribosomal RNA (rRNA) gene sequences from intestinal biopsies as well as host genotype via Immunochip in three independent cohorts totaling 474 individuals. We tested for correlation between relative abundance of bacterial taxa and number of minor alleles at known IBD risk loci, including fine mapping of multiple risk alleles in the Nucleotide-binding oligomerization domain-containing protein 2 (*NOD2*) gene exon. We identified host polymorphisms whose associations with bacterial taxa were conserved across two or more cohorts, and we tested related genes for enrichment of host functional pathways.

**Results:** We identified and confirmed in two cohorts a significant association between *NOD2* risk allele count and increased relative abundance of Enterobacteriaceae, with directionality of the effect conserved in the third cohort. Forty-eight additional IBD-related SNPs have directionality of their associations with bacterial taxa significantly conserved across two or three cohorts, implicating genes enriched for regulation of innate immune response, the JAK-STAT cascade, and other immunity-related pathways.

**Conclusions:** These results suggest complex interactions between genetically altered host functional pathways and the structure of the microbiome. Our findings demonstrate the ability to uncover novel associations from paired genome-microbiome data, and they suggest a complex link between host genetics and microbial dysbiosis in subjects with IBD across independent cohorts.

* Correspondence: dknights@umn.edu; xavier@molbio.mgh.harvard.edu
†Equal contributors
[1]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, Minnesota 55455, USA
[2]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA
Full list of author information is available at the end of the article

## Background

Crohn's disease (CD) and ulcerative colitis (UC), collectively known as inflammatory bowel disease (IBD), have long been known to have genetic risk factors due to increased prevalence in relatives of affected individuals as well as higher concordance rates for disease among monozygotic versus dizygotic twins. The sequencing of the human genome and subsequent large-cohort genetic studies has revealed a complex set of polymorphisms conferring varying levels of risk. Extensive analyses of these loci revealed that impaired handling of commensal microbes and pathogens is a prominent factor in disease development [1]. For example, genetically driven impaired function of NOD2 in the sensing of bacterial products like lipopolysaccharide may cause an increase in bacteria that produce those products. Involvement of the JAK-STAT pathway in immune responses, and involvement of the IL-23-Th17 pathway in microbial defense mechanisms, are also possible links between impaired immune response and imbalances in bacterial assemblage [1-3]. These genetic findings are in line with separate, independent tests of microbial shifts associated with IBD. Shifts in taxonomic composition and metabolic capabilities of the IBD microbiome are both now beginning to be defined [4-9]. Determining the extent and nature of host genome-microbiome associations in IBD is an important next step in understanding the mechanisms of pathogenesis. Despite the documented independent associations of IBD with heritable host immune deficiencies and with microbial shifts, there has been limited study of the co-association of complex host genetic factors with microbial composition and metabolism in IBD patients or other populations [9-17], and the mechanisms of host-microbiome disease pathways are largely unknown.

Using three independent cohorts comprising 474 adult human subjects with IBD aged 18 to 75 years, we tested known IBD-associated host genetic loci for enrichment of association with gut microbiome taxonomic composition. Cohorts were located near Boston (USA), Toronto (Canada), and Groningen (the Netherlands), with 152, 160, and 162 subjects, respectively. The cohorts contained 62.5%, 14.3%, and 63.5% CD cases with the remainder cases of UC, and 31.5%, 11.3%, and 53.1% biopsies from inflamed sites, respectively (detailed summary statistics by cohort and biopsy location in Figures S1 and S2 in Additional file 1). The Toronto cohort contained 70.6% biopsies from the pre-pouch ileum in subjects with previous ileo-anal pouch surgery; all remaining samples were from the colon and terminal ileum, with 73.0%, 18.1%, and 87.0% from the colon in the three cohorts, respectively. We excluded all subjects that had taken antibiotics within one month prior to sampling. We obtained genotyping with Illumina Immunochip assays [18] and 16S rRNA gene sequences as described previously [19] (SNP prevalence by cohort in Additional file 2). We rarefied bacterial microbiome samples to an even sequencing depth of 2,000 sequences per sample to control for differential sequencing effort across cohorts. This rarefaction depth allows us to observe taxa with relative abundance as low as 0.15% with 95% confidence in each sample (binomial distribution with 2,000 trials and probability 0.0015). We report a pathway-level analysis of complex functional associations between host genetics and overall microbiome composition, as well as a targeted analysis of the association of *NOD2* with specific bacterial taxa.

## Methods

### Ethics and consent

This study was approved by the Partners Human Research Committee, 116 Huntington Avenue, Boston, MA, USA. Patients gave informed consent to participate in the study. This study conformed to the Helsinki Declaration and to local legislation.

### Data collection and generation

We genotyped subjects using the Immunochip platform as described previously [18], excluding polymorphisms with minor allele frequency of 0.1 or below from subsequent testing. 16S rRNA genes were extracted and amplified from intestinal biopsies and sequenced on the Illumina MiSeq platform using published methods [20]. These procedures include extraction using the QIAamp DNA Stool Mini Kit (Qiagen, Inc., Valencia, CA, USA) according to the manufacturer's instructions with minor alterations described in prior work [20], followed by amplification using the 16S variable region 4 forward primer GTGCCAGCMGCCGCGGTAA and reverse primer GGACTACHVGGGTWTCTAAT, followed by barcoded multiplexing and sequencing. Only one biopsy was used per subject; when multiple biopsies were available we selected the non-inflamed biopsy first.

### Data processing

We extracted risk allele counts for 163 published genetic risk loci for CD, UC, and IBD [1]. When combining data from separate Immunochip runs we tested for strand inversions by linkage disequilibrium with neighboring variants using plink [21]. Microbial operational taxonomic units (OTUs) and their taxonomic assignments were obtained using default settings in QIIME version 1.8 [22] by reference-mapping at 97% similarity against representative sequences of 97% OTU in Greengenes (taxa version 4feb2011; metagenome version 12_10) [23]. We used all default settings in QIIME 1.8 for OTU mapping, and we used the pre-assigned taxonomy for the Greengenes OTU representative sequences. Samples were rarefied to an even sequence depth of 2,000 sequences per sample to control for varied sequencing depth. Taxa were collapsed into clusters with >0.95 Pearson's correlation to remove

redundant signals in the data (Additional file 3). Principal coordinates of between-subject distances were obtained from UniFrac [24] distances of OTUs and Jensen-Shannon and Bray-Curtis distances of KEGG (Kyoto Encyclopedia of Genes and Genomes) module and pathway distributions. Bacterial taxa were arcsine-square-root transformed and bacterial functions were power-transformed ('car' package [25]) to stabilize variance and reduce heteroscedasticity.

### Statistical analysis

Linear association tests were performed only within those taxa with nonzero abundance in at least 75% of subjects. Taxa below that threshold were subjected to logistic regression for presence/absence; no such taxa revealed significant associations after correcting for multiple comparisons. To ensure robustness of tests to outliers, subjects with taxon or functional module relative abundance more than three times the interquartile range from the mean were excluded for tests of that feature. Power analysis was performed using the linear effect size that we observed for Enterobacteriaceae when regressing on *NOD2* risk allele count and controlling linearly for clinical covariates ($f^2 = R^2/(1 - R^2) = 0.013$; R is the coefficient of multiple correlation). Assuming the need to correct for testing of all 163 IBD loci against 22 dominant taxa (3,586 tests; adjusted significance threshold = $1.39 \times 10^{-5}$), we would need at least 3,729 samples to power the full analysis (R 'pwr' package power calculation for a linear model with 19 numerator degrees of freedom). Discrete qualitative covariates were re-coded with dichotomous dummy variables representing each class prior to testing. Association of clinical covariates was performed jointly by multiple linear regression. To overcome redundancy between clinical covariates, we clustered clinical covariates based on their pairwise maximum uncertainty coefficients [26], an information-theoretic measure of their degrees of shared information. Continuous-valued covariates were discretized prior to information-theoretic clustering. Complete-linkage clustering was performed to identify groups of covariates in which each covariate contained at least 50% of the information contained in each other covariate. Network plots were created using the igraph [27] package. For the network plot of non-genetic host factors and *NOD2*, width of edges was determined by the ratio of a given covariate's linear regression coefficient to the mean of the regressed taxon's relative abundance. Enrichment of a host functional pathway for association with bacterial taxa was assessed by comparing the observed rank product of all host gene-bacterial taxon association tests for all genes in the pathway with the distribution of rank products of 100,000 size-matched pathways randomly generated from the null Immunochip variants described above. Prior to testing, REACTOME pathways with >75% overlap were binned and the largest constituent pathway chosen as a representative for subsequent tests.

## Results and discussion

### Genotype-microbiome associations conserved across independent cohorts

Our genotype-microbiome association testing methodology included steps to overcome power limitations given the very large number of potential comparisons, to incorporate published knowledge of signaling and metabolic pathways in the host genome, and to control for multiple environmental host factors affecting gut microbiome composition (Figure 1). In a targeted analysis of *NOD2*, we also accounted for multiple causal variants in the genetic locus (Supplementary methods in Additional file 1). After data preprocessing and normalization we tested linearly for association of risk allele count in each SNP with the relative abundance of each bacterial taxon. In all tests, we controlled for recent antibiotic usage (<1 month), recent immunosuppressant usage (<1 month), biopsy inflammation status based on pathology, age, gender, biopsy location, CD/UC diagnosis, disease location, elapsed time since diagnosis, cohort membership, and the first three principal components of genotype variation (Figure 1; Figure S3 in Additional file 1). Although the IBD-related SNPs extracted from the Immunochip data were identified previously in European populations, we do not expect this to limit our findings because our cohorts were mostly of European descent. We validated our linear testing methodology by comparing associations in the Boston cohort with those in the other two cohorts, in addition to performing other sensitivity analyses (Supplementary methods in Additional file 1).

We tested 163 recently IBD-associated SNPs for association with bacterial taxonomic profiles; 154 remained after removing those with low minor allele frequencies or with low call rates in our cohorts (Supplementary methods in Additional file 1). Many of these SNPs have unknown mechanisms and are likely only representative of a signal within the surrounding genomic locus. Therefore, when a single gene was associated previously with a SNP, we refer to that SNP by the gene name for convenience. Due to limited statistical power we were unable to perform a full analysis of all possible SNP-taxon associations (Supplementary methods in Additional file 1). However, we were able to test for the robustness of microbiome-wide associations with a given SNP by comparing the directionality of the SNP-taxon coefficients between independent cohorts. For this test we included only those SNP-taxon associations for a given SNP that were nominally significant ($P < 0.05$) in at least one of the studies being compared. We then obtained Matthew's correlation coefficient (MCC; also known as the phi coefficient) of the signs (positive or negative) of the SNP-taxon coefficients in one study with the signs of corresponding SNP-taxon coefficients in the second study, and corrected these microbiome-wide tests for multiple comparisons (one MCC test per gene) at a false discovery rate (FDR) of 0.25.

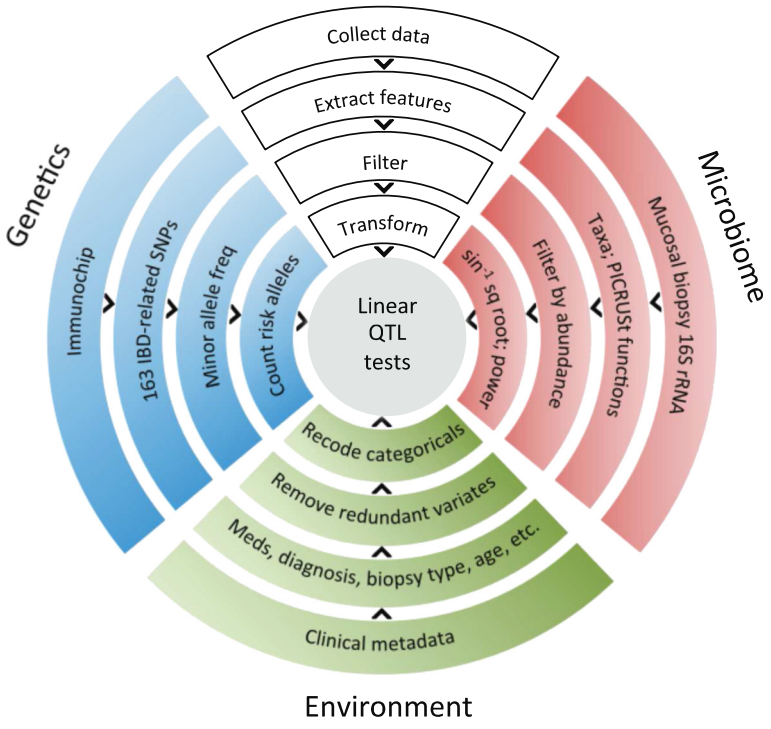


**Figure 1 Schematic of multiomics genotype-microbiome association testing methodology.** Host genome-microbiome association testing involves potentially thousands or millions of genetic polymorphisms and hundreds or thousands of bacterial taxa and genes. Full feature-by-feature association testing is likely to be underpowered in all but the largest cohorts or meta-analyses; therefore, our methodology includes careful feature selection from both data types. Raw genetic polymorphisms were derived from Immunochip data and filtered by known IBD associations from a large-cohort GWAS study [1]. Microbiome sequences were binned by lineage at all taxonomic levels. After data normalization and filtering (see Methods), a simple linear test was performed for association between minor allele count and bacterial taxon relative abundance while controlling for clinical covariates. QTL, quantitative trait loci.

We chose the FDR of 0.25 for this analysis due to the large number of tests and the fact that we used the significant results mainly to test for enrichment of certain host pathways, rather than to focus on individual associations. We note that it is important to compare only the directionality of SNP-taxon effects between studies, and not the magnitudes of the SNP-taxon regression coefficients, because the magnitude of a coefficient is closely linked to the mean relative abundance of a given taxon. To decrease bias toward a particular taxonomic level of association [28], we performed these tests using bacterial taxa at all taxonomic levels from phylum to genus, collapsing those with redundant signals. In contrast to using OTU clusters, binning by taxonomy allows inherent flexibility in the level of 16S gene sequence identity within each bin in different lineages.

A number of host genes, some with known involvement in microbial handling, and others with unknown function, demonstrated reproducible effects on the taxonomic structure of the microbiome across two or more cohorts. Effect size and directionality of genotype-microbiome associations were highly reproducible between cohorts in the case of *NOD2* and 48 other host genes (FDR <0.25; Additional file 4). *NOD2* had one of the most highly reproducible sets of associations with bacterial taxa (MCC = 0.75, FDR = $2.6 \times 10^{-4}$ comparing Boston versus Toronto cohorts; MCC = 0.85, FDR = $7.7 \times 10^{-4}$ Boston versus Netherlands; Figure 2a). Other genes with significantly conserved directionality of effects on bacterial taxa between at least one pair of studies included tumor necrosis factor (ligand) superfamily, member 15 (*TNFSF15*; MCC = 0.87, FDR = $9.5 \times 10^{-3}$, Boston versus Netherlands) and subunit beta of interleukin 12 (*IL12B*; MCC = 0.74, FDR = $1.5 \times 10^{-3}$, Boston versus Netherlands).

*NOD2* variants were the first genetic associations identified in CD, and they remain some of the strongest risk factors. *NOD2*-driven murine dysbiosis causes inflammation even when the dysbiotic microbiota are transplanted into a wild-type mouse [13]. Expression of *TNFSF15*, a member of the tumor necrosis factor ligand superfamily, causes proinflammatory cytokine production, and is specifically expressed more highly in the gut in IBD patients compared with healthy controls. Interestingly, a receptor for a member of the same family, TNFSF14, enhances immune response to pathogenic bacteria via signal transducer and activator of transcription 3 (*STAT3*) activation

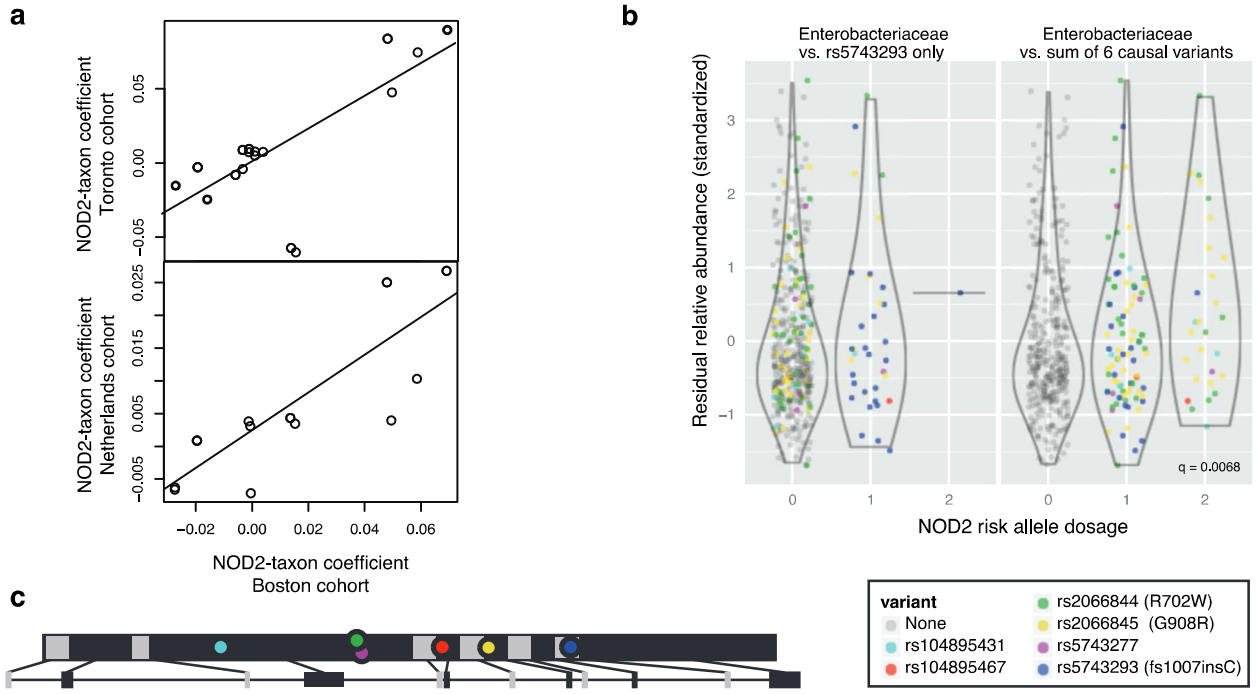


**Figure 2 *NOD2* fine mapping reveals association with taxonomic and metabolic dysbiosis. (a)** Scatterplot of NOD2-bacterial taxon regression coefficients in one study versus the corresponding regression coefficients in another study. We included only those taxa with a nominally significant ($P < 0.05$) association in a least one of the cohorts being compared. **(b)** Comparison of residual distributions of Enterobacteriaceae with and without incorporating the six independent known causal *NOD2* variants; considering variant rs5743293, only 6.3% of subjects have one or more risk alleles; aggregating risk allele counts across the six variants increases this to 21.8%, and reveals much stronger associations with the microbiome. The strip charts and violin plots show the distribution of standardized residual relative abundance of Enterobacteriaceae versus *NOD2* risk allele dosage after data transformation and regression on clinical covariates. Violin plots show the conditional density of residual relative abundance within each dosage level. **(c)** Relative positions of six *NOD2* variants in *NOD2* exons [29].

in a mouse model of *Escherichia coli* infection. TNFSF14 and TNFSF15 are known to share an alternative receptor, indicating potential functional overlap. IL12B forms part of the interleukin-23 complex, involved in microbial defense mechanisms through the IL23-Th17 pathway.

### Immunity-related host functional pathways linked to microbiome profile

We hypothesized that host functional pathways containing multiple risk variants related to microbial handling and innate immune response would be associated with microbiome features. To test this hypothesis we performed a functional enrichment analysis on the 49 genes identified to have conserved microbiome associations across cohorts. We found these genes to be significantly enriched for regulation of innate immune response (FDR = $2.31 \times 10^{-6}$, hypergeometric enrichment test), inflammatory response (FDR = $7.43 \times 10^{-6}$, hypergeometric enrichment test), and participation in the JAK-STAT cascade (FDR = $2.04 \times 10^{-4}$, hypergeometric enrichment test) (Figures 3 and 4; Additional file 5). A gene-gene interaction network analysis also implicated *STAT3*, interleukin-12 subunit alpha (*IL12A*), and interleukin-23 subunit alpha (*IL23A*) in the network of associated genes.

*STAT3* and *TNFSF15* are both implicated in IL23 signaling. *STAT3* works in concert with Janus Kinase 2 (*JAK2*) in the JAK-STAT pathway to drive immune response to pathogenic infection. *STAT3* also regulates T helper 17 (Th17) cell differentiation by binding IL23 receptor (IL23R; risk variant for IBD: rs11209026) and RAR-related orphan receptor C (RORC; rs4845604), both of which are located in IBD risk loci. *STAT3* defects have also been implicated recently in skin microbial imbalance and impaired host defense. TNFSF15, a member of the tumor necrosis factor ligand superfamily, is a costimulator of T cells, and is specifically expressed more highly in the gut in IBD patients compared with healthy controls [31,32].

### Fine mapping of *NOD2* locus reveals association with Enterobacteriaceae

Based on previous results [9-13] and on the strong linkage between NOD2 and microbial handling [9,12,13], we continued with a targeted analysis of *NOD2* association with specific microbial taxa and functions (Additional file 6).

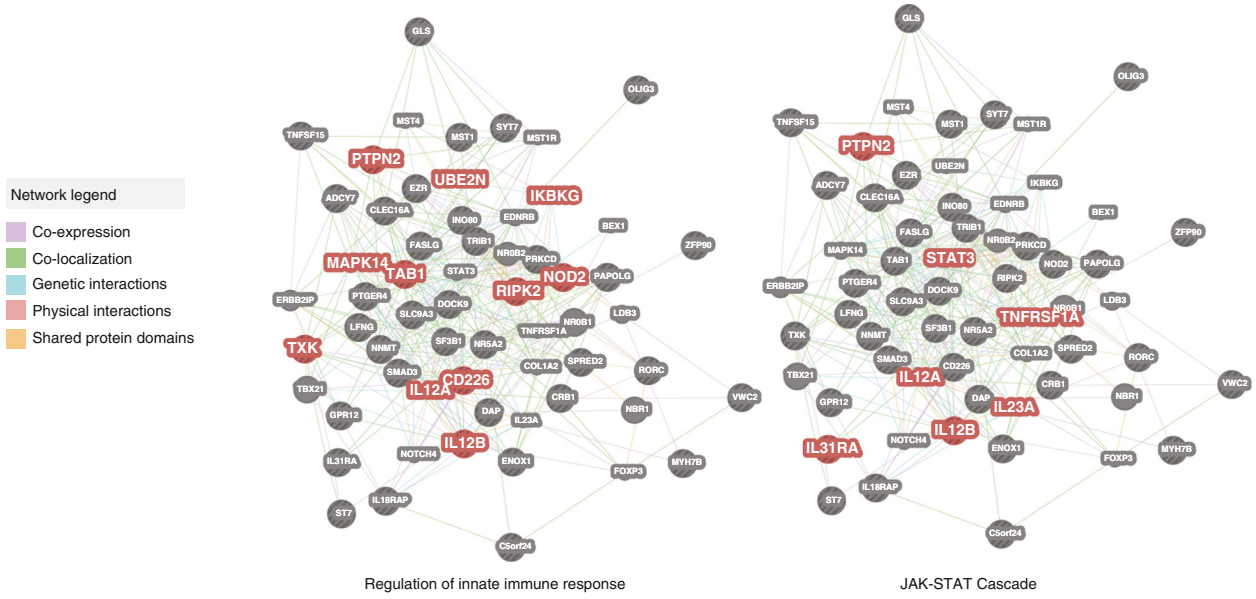


**Figure 3 Host genes with reproducible microbiome associations across cohorts.** Network analysis of host signaling and metabolic pathways enriched for association with microbial taxa (FDR <0.25, Matthew's correlation test). The visualization of gene-gene interaction network for the subset of 49 genes with significantly conserved directionalities of association with the microbiome is supported by several types of gene-gene connection [30]. This enrichment analysis identified enriched functional networks in innate immune response, inflammatory response, and the JAK-STAT pathway, all of which play roles in immune response to pathogen infection [1].

For all *NOD2* testing we aggregated risk allele dosage across six known causal variants: rs2066844 (R702W), rs2066845 (G908R), rs5743277, rs5743293 (fs1007insC), rs104895431 and rs104895467 [29]. This novel methodology was crucial as individual variants contained only part of the signal (Figure 2b,c). *NOD2* was associated with the first principal axis of taxonomic (via weighted UniFrac distances) microbiome variation (FDR <0.05, controlling for three principal axes tested), linking *NOD2* to shifts in overall microbiome taxonomic composition. We identified increased Enterobacteriaceae in subjects with higher *NOD2* risk allele dosage (FDR = 0.11, controlling for multiple taxa tested; Figure 2b; Additional file 7). An increase in Gammaproteobacteria is a known component of IBD dysbiosis and is associated with inflammation in mice and humans [4,33] and with increased epithelial penetration in CD and UC [34]. *NOD2* also had one of the most strongly reproducible associations with microbiome composition when comparing cohorts. Although *NOD2* is only associated with increased risk of CD, *NOD2*-microbiome associations we observed were generally independent of CD/UC diagnosis, with high overlap between CD and UC when tested separately (taxa: Spearman's rho = 0.57, $P = 6.5 \times 10^{-3}$; Figure S4 in Additional file 1). This implies that the association may be disease-independent, and may play a role in pathogenesis only in subjects with other risk factors. For example, *NOD2* SNP rs5743293 is associated with complications in ileo-anal pouch patients despite their original diagnosis being UC [35-38].

## A complex web of genotype-environment-microbiome interactions in IBD

Our findings indicate that host genetics are part of a complex web of host-associated factors influencing microbiome composition. We performed a meta-analysis of interactions between clinical host factors and bacterial taxa using the above 474 subjects and an additional 55 subjects who had recently taken antibiotics. This analysis included *NOD2* as one of the host factors. We identified an additional 99 significant associations of non-genetic factors with relative abundance of specific bacterial taxa, largely in agreement with previous analyses of the IBD microbial environment [4]. To visualize the overlap of interactions between various host factors and microbial taxa, we constructed a network of associations between bacterial taxa and observed host and environmental factors (Figure 5; Additional file 8; Supplementary methods in Additional file 1).

This analysis revealed a web of complex overlapping linkages between numerous host factors and bacterial taxa. For example, recent antibiotic usage is associated with systematic shifts in many major taxonomic groups (Figure 6; FDR <0.05); immunosuppressants are associated with decreased Firmicutes, and Ruminococcaceae. Biopsy location and

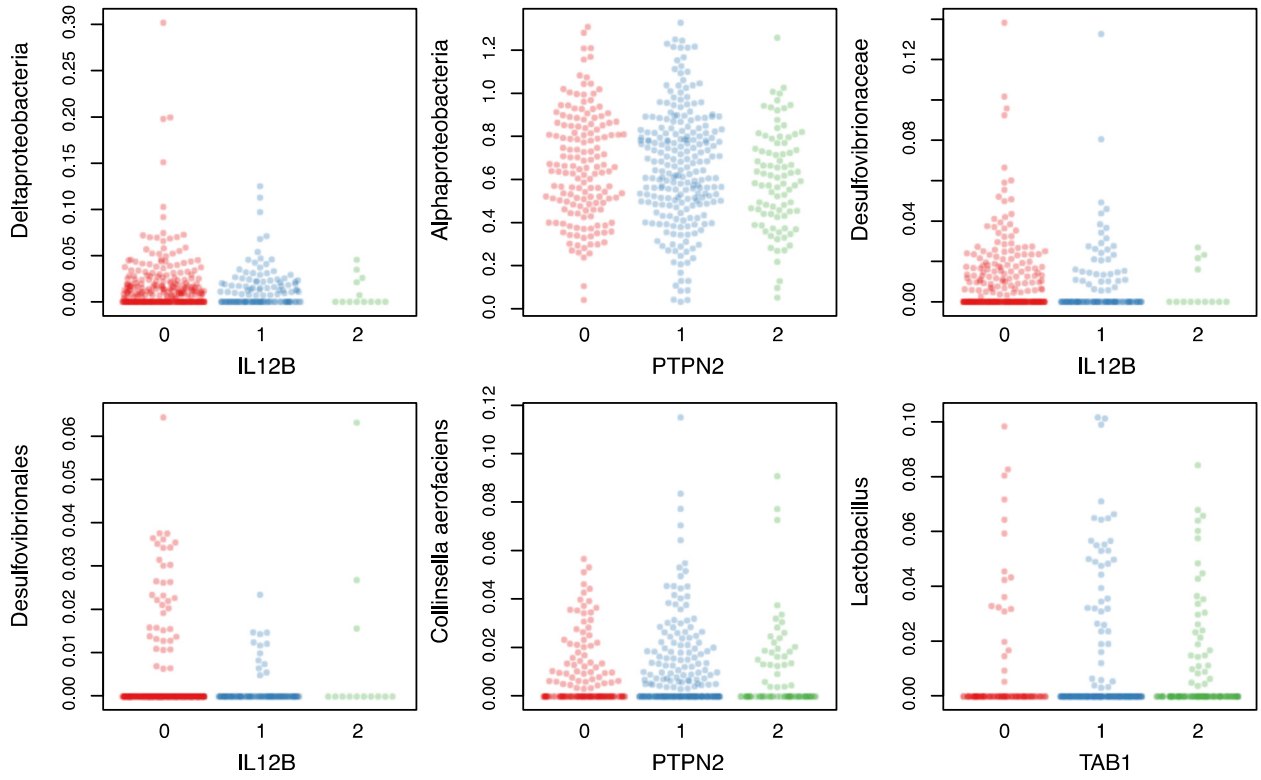


**Figure 4 Top gene-bacteria associations.** Beeswarm plots of the relative abundance of six bacteria stratified by the number of risk alleles present in SNPs in the given genes. The associations shown are the six most significant associations between bacteria and genes in the subset of genes with conserved bacterial associations across cohorts and belonging to the JAK-STAT pathway or the innate immune pathway response as shown in Figure 3. Relative abundances shown are transformed with the arcsine-square root transformation to stabilize variance and to make distributions more normal.

cohort membership had similarly broad effects; age, gender, and disease phenotype had measurable, although less broad, effects; genotype, as represented by the *NOD2* subtype, had a modest effect in relation to other factors. Inflammatory status of the biopsied tissue was associated with increased relative abundance of unclassified members of *Lactobacillus*, and with decreased relative abundance of *Bacteroides uniformis* (Figure S5 in Additional file 1). This analysis demonstrates the comprehensive and intermingled effects of treatment history, gastrointestinal biogeography, and other host and environmental factors on gut microbiome profile and makes clear the need to account for host factors when linking host genotype to microbial composition in a phenotypically heterogeneous population. We confirmed that host genetics as a whole do have a significant effect on microbiome profile by correlating overall between-subject genetic distance (Manhattan distance) with overall between-subject microbiome distance (unweighted UniFrac distance) ($P < 5.0 \times 10^{-10}$; Figure S6 in Additional file 1), but that it is only a minor contributor in the context of other sources of variation. A recent study of treatment-naïve pediatric patients with CD identified consistent microbiome shifts in patients with recent antibiotic exposure toward the disease-related state [20]. That study exemplified the need to control for the potentially confounding effects of antibiotics when attempting to identify bacterial profiles associated with disease. Based on several studies linking short- and long-term dietary exposure to microbiome profile, it is also likely to be useful to include food intake diaries or dietary recall questionnaires in future genotype-microbiome research [39,40].

### Antibiotics contribute to IBD dysbiosis independent of NOD2 effects

The fact that host genetics are a minor contributor to overall microbiome composition relative to environmental factors does not exclude the possibility that genotype-microbiome interactions play an important role in the etiology of IBD; it is possible that the important variations are in a particular set of taxa or a particular set of functions (for example, resistance to oxidative stress) that make up a minor portion of the overall microbiome, while there are other taxa not closely related to IBD but highly influenced by the host's environmental exposures (for example, dietary exposures). Such a subset of taxa related to dysbiosis were reported in a recent comparison between treatment-naïve patients with Crohn's disease and healthy controls [20], and the ratio of the disease-associated taxa to the health-associated taxa was referred to as the

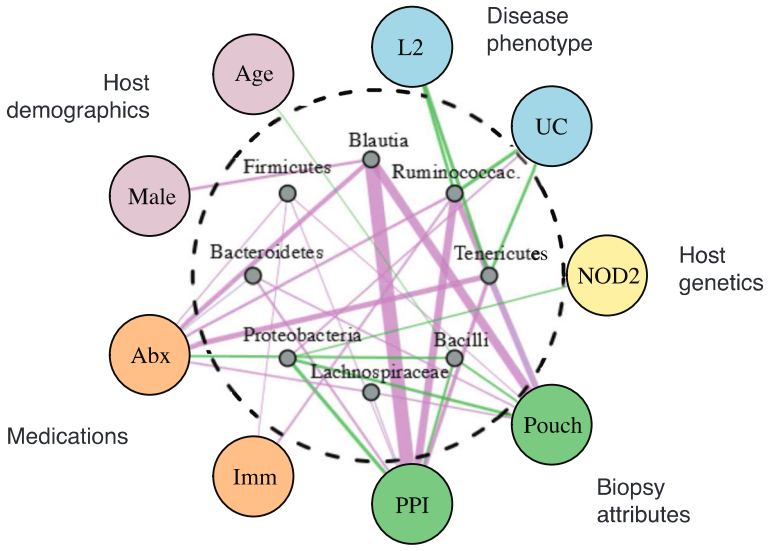


**Figure 5 Host factors associated with the IBD microbiome.** A complex network of host factors associated with the IBD microbiome (all associations FDR <0.05); only taxa with at least four significant associations are included in the network; green and purple edges indicate positive and negative associations, respectively; the width of an edge indicates the strength of the association. The effects of these factors on individual taxa are highly overlapping. The analysis identified covariates representing each type of host factor, consistent with previous results [4]. Biopsy location and medication history had the strongest and most comprehensive effects on microbiome profile; the effect of *NOD2* was moderate in comparison. Cohort membership (not shown) also affected microbiome profile. These results demonstrate the need for study designs and analysis methodologies that control carefully for numerous host genetic and environmental factors when performing microbiome-based biomarker discovery. Abx, antibiotics within 1 month; Imm, immunosuppressants within 1 month; L2, no ileal involvement; PPI, biopsy from pre-pouch ileum.

microbial dysbiosis index (MDI). This recent study identified an increase in the MDI scores of patients who had recently received antibiotics, indicating that antibiotics tend to shift patient microbiomes further into the realm of IBD-related dysbiosis. We used the same taxa as reported previously to calculate an MDI score for each patient in our analysis. In our cohorts we confirmed the published finding that, when controlling for NOD2 effects on microbiome structure, MDI score tended to be higher in patients with recent usage (within less than one month) of antibiotics ($P = 0.039$, $t$-test of linear regression coefficient) (Figure 6). This finding, together with previously published findings regarding the effects of antibiotics on the IBD microbiome suggest that antibiotics and duration of disease are additional risk factors for IBD-related dysbiosis.

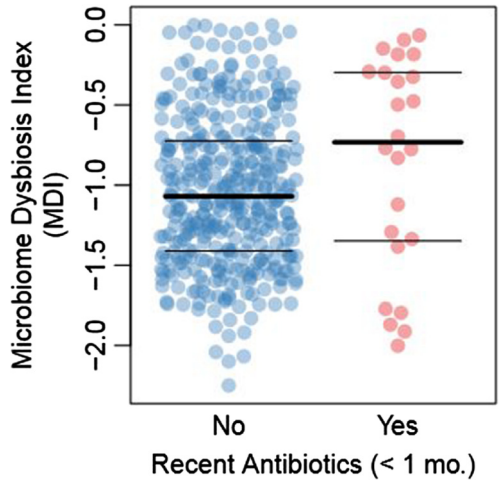


**Figure 6 Association of IBD-related dysbiosis and recent antibiotics usage.** A beeswarm plot [41] of the previously published microbial dysbiosis index [20] (MDI) stratified by recent antibiotics usage by patients. The test for this association between MDI and antibiotics ($P = 0.039$, linear regression $t$-test) included *NOD2* risk allele count to control for the effects of *NOD2* genetics on the microbiome.

## Conclusions

Taken together, our findings indicate a complex set of associations between the mucosal-adherent microbiome and genetic impairment of several host immune pathways. Although we have been living and evolving with our microbial symbionts throughout human evolution, we have only been aware of their existence for a few centuries, and the genetic and functional diversity of our so-called 'second genome' has only become apparent in the last few decades. Also in recent decades incidence of IBDs and other autoimmune and autoinflammatory diseases has increased dramatically [42], and a rapidly growing set of these diseases has been linked to shifts both in taxonomic carriage and functional potential of host-associated microbial communities. Although our data are cross-sectional and therefore cannot define

causality, our analyses demonstrate complex host genetic associations with taxonomic and metabolic dysbiosis in humans. These include implications of microbiome-wide associations with *TNFSF15*, *IL12B*, and with innate immune response, inflammatory response, and the JAK-STAT pathway, as well as *NOD2*-related increases in Enterobacteriaceae relative abundance. Future studies may be warranted to account for the effects of copy number variation, pleiotropic genes and epigenetic modifications. It is also possible that certain genotype-microbiome associations observed in IBD patients may be disease-independent and may be relevant to healthy individuals and individuals with other diseases. The methods we employed were validated on independent cohorts and make possible well-powered false-positive-controlled testing of microbiome-wide host genetic associations.

### Accession numbers

16S rRNA sequences and Immunochip genotyping have been deposited at the National Center for Biotechnology Information as BioProject with top-level umbrella project ID PRJNA205152.

## Additional files

**Additional file 1: Supplementary methods and figures.**

**Additional file 2: Table S1.** The prevalence of 163 known IBD-associated SNPs in the three independent cohorts included in this study.

**Additional file 3: Table S2.** Groups of taxa that were binned together prior to statistical testing due to high inter-group correlation.

**Additional file 4: Table S3.** Correlations and significance levels thereof of the directionalities of SNP-taxon associations between pairs of cohorts for those associations that were nominally significant in at least one of the cohorts being compared.

**Additional file 5: Table S4** Host genetic functional pathways that were significantly enriched for robustness of association with the microbiome across cohorts.

**Additional file 6: Specific SNPs that were included in gene-level fine-mapping of disease association signals.**

**Additional file 7: Linear association test results for the association of each IBD SNP with each bacterial taxon.**

**Additional file 8: Additional association test results for a meta-analysis of the three cohorts, including statistics for the associations of clinical covariates with bacterial taxa.**

#### Abbreviations

CD: Crohn's disease; FDR: false discovery rate; IBD: inflammatory bowel disease; IL: interleukin; MCC: Matthew's correlation coefficient; MDI: microbial dysbiosis index; OTU: operational taxonomic unit; SNP: single nucleotide polymorphism; Th17: T helper 17; UC: ulcerative colitis.

#### Competing interests

The authors declare that they have no competing interests.

#### Authors' contributions

DK, MSS, RKW, and RJX wrote the manuscript; JK, JS, AT, DG, and CH reviewed and revised the manuscript; MSS, GD, RKW, and RJX established biopsy and DNA collections; DK, JK, MSS, RKW, SvS, and Hailing H performed or supervised genotyping quality control and statistical analysis; DK and DG performed microbiome quality control and statistical analysis; AT performed pouch sample collection and phenotyping; DK, PV, GA, and CH developed genome-microbiome association methods and performed association tests; Hu H analyzed shotgun metagenomic data; CH and RJX supervised methods development and statistical analysis; FI, JS and CR collected clinical data and managed the clinical databases; CR performed microbiome sample extraction. All authors read and approved the final manuscript.

#### Acknowledgements

We thank the patients who donated samples for this study, and the health professionals who collected them. We thank Tjasso Blokzijl for technical assistance and Levi Waldron for power analysis code. We thank Aylwin Ng and Moran Yassour for critical review of the manuscript. We thank Timothy Tickle and Tonya Ward for helpful discussions regarding methods. RKW is supported by a VIDI grant from the Netherlands Organization for Scientific Research (NWO) and the Dutch Digestive Foundation (WO 11-72). CH is partially supported by NIH R01HG005969, NSF DBI-1053486, and ARO W911NF-11-1-0473. MSS is partially supported by the Gale and Graham Wright Research Chair in Digestive Disease. Partial funding for sample collection for Toronto samples provided by the Crohn's and Colitis Foundation of Canada. Work was supported by grants from the Crohn's and Colitis Foundation of America, NIH grants U54 DE023798, and R01 DK092405 (RJX, CH, DG). JK is the Joanne Murphy Professor in Behavioural Science.

#### Author details

[1]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, Minnesota 55455, USA. [2]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. [3]Center for Computational and Integrative Biology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA. [4]Biotechnology Institute, University of Minnesota, St. Paul, Minnesota 55108, USA. [5]Zane Cohen Centre for Digestive Diseases, Mount Sinai Hospital IBD Group, University of Toronto, Toronto, Ontario M5G 1X5, Canada. [6]Department of Gastroenterology and Hepatology, University Medical Center Groningen, Groningen 9700RB, The Netherlands. [7]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [8]Department of Genetics, University Medical Center Groningen, Groningen 9700RB, The Netherlands. [9]Biomedical Informatics and Computational Biology, University of Minnesota, Minneapolis, Minnesota 55455, USA. [10]Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA. [11]Department of Psychiatry, University of Toronto, Toronto, Ontario M5T 1R8, Canada. [12]Department of Medicine, Analytic and Translational Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA. [13]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. [14]Biostatistics Department, Harvard School of Public Health, Boston, Massachusetts 02115, USA.

Received: 2 September 2014 Accepted: 13 November 2014
Published online: 02 December 2014

#### References

1. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, Lee JC, Schumm LP, Sharma Y, Anderson CA, Essers J, Mitrovic M, Ning K, Cleynen I, Theatre E, Spain SL, Raychaudhuri S, Goyette P, Wei Z, Abraham C, Achkar JP, Ahmad T, Amininejad L, Ananthakrishnan AN, Andersen V, Andrews JM, Baidoo L, Balschun T, Bampton PA, Bitton A, *et al*: **Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease.** *Nature* 2012, **491:**119–124.
2. Abraham C, Cho J: **Interleukin-23/Th17 pathways and inflammatory bowel disease.** *Inflamm Bowel Dis* 2009, **15:**1090–1100.
3. Knights D, Lassen KG, Xavier RJ: **Advances in inflammatory bowel disease pathogenesis: linking host genetics and the microbiome.** *Gut* 2013, **62:**1505–1510.
4. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, Reyes JA, Shah SA, LeLeiko N, Snapper SB, Bousvaros A, Korzenik J, Sands BE, Xavier RJ, Huttenhower C: **Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment.** *Genome Biol* 2012, **13:**R79.
5. Baumgart M, Dogan B, Rishniw M, Weitzman G, Bosworth B, Yantiss R, Orsi RH, Wiedmann M, McDonough P, Kim SG, Berg D, Schukken Y, Scherl E, Simpson KW: **Culture independent analysis of ileal mucosa reveals a selective increase in**

invasive Escherichia coli of novel phylogeny relative to depletion of Clostridiales in Crohn's disease involving the ileum. *ISME J* 2007, **1:**403–418.
6. Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR: **Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases.** *Proc Natl Acad Sci U S A* 2007, **104:**13780–13785.
7. Scanlan PD, Shanahan F, O'Mahony C, Marchesi JR: **Culture-independent analyses of temporal variation of the dominant fecal microbiota and targeted bacterial subgroups in Crohn's disease.** *J Clin Microbiol* 2006, **44:**3980–3988.
8. Li E, Hamm CM, Gulati AS, Sartor RB, Chen H, Wu X, Zhang T, Rohlf FJ, Zhu W, Gu C, Robertson CE, Pace NR, Boedeker EC, Harpaz N, Yuan J, Weinstock GM, Sodergren E, Frank DN: **Inflammatory bowel diseases phenotype, C. difficile and NOD2 genotype are associated with shifts in human ileum associated microbial composition.** *PLoS One* 2012, **7:**e26284.
9. Dalton JP, Desmond A, Shanahan F, Hall C: **Detection of Mycobacterium avium subspecies paratuberculosis in patients with Crohn's disease is unrelated to the presence of single nucleotide polymorphisms rs2241880 (ATG16L1) and rs10045431 (IL12B).** *Med microbiol and immun* 2014, **203:**195–205.
10. Ott SJ, Musfeldt M, Wenderoth DF, Hampe J, Brant O, Fölsch UR, Timmis KN, Schreiber S: **Reduction in diversity of the colonic mucosa associated bacterial microflora in patients with active inflammatory bowel disease.** *Gut* 2004, **53:**685–693.
11. Smeekens SP, Huttenhower C, Riza A, van de Veerdonk FL, Zeeuwen PLJM, Schalkwijk J, van der Meer JWM, Xavier RJ, Netea MG, Gevers D: **Skin microbiome imbalance in patients with STAT1/STAT3 defects impairs innate host defense responses.** *J Innate Immun* 2013, **6:**253–262.
12. Frank DN, Robertson CE, Hamm CM, Kpadeh Z, Zhang T, Chen H, Zhu W, Sartor RB, Boedeker EC, Harpaz N, Pace NR, Li E: **Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases.** *Inflamm Bowel Dis* 2011, **17:**179–184.
13. Couturier-Maillard A, Secher T, Rehman A, Normand S, De Arcangelis A, Haesler R, Huot L, Grandjean T, Bressenot A, Delanoye-Crespin A, Gaillot O, Schreiber S, Lemoine Y, Ryffel B, Hot D, Nùñez G, Chen G, Rosenstiel P, Chamaillard M: **NOD2-mediated dysbiosis predisposes mice to transmissible colitis and colorectal cancer.** *J Clin Invest* 2013, **123:**700–711.
14. Rajilić-Stojanović M, Smidt H, de Vos WM: **Diversity of the human gastrointestinal tract microbiota revisited.** *Environ Microbiol* 2007, **9:**2125–2136.
15. Hansen EE, Lozupone CA, Rey FE, Wu M, Guruge JL, Narra A, Goodfellow J, Zaneveld JR, McDonald DT, Goodrich JA, Heath AC, Knight R, Gordon JI: **Pan-genome of the dominant human gut-associated archaeon, Methanobrevibacter smithii, studied in twins.** *Proc Natl Acad Sci U S A* 2011, **108:**4599–4606.
16. Ley RE, Bäckhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI: **Obesity alters gut microbial ecology.** *Proc Natl Acad Sci U S A* 2005, **102:**11070–11075.
17. Hashimoto T, Perlot T, Rehman A, Trichereau J, Ishiguro H, Paolino M, Sigl V, Hanada T, Hanada R, Lipinski S, Wild B, Camargo SM, Singer D, Richter A, Kuba K, Fukamizu A, Schreiber S, Clevers H, Verrey F, Rosenstiel P, Penninger JM: **ACE2 links amino acid malnutrition to microbial ecology and intestinal inflammation.** *Nature* 2012, **487:**477–481.
18. Cortes A, Brown MA: **Promise and pitfalls of the Immunochip.** *Arthritis Res Ther* 2011, **13:**101.
19. Methé BA, Nelson KE, Pop M, Creasy HH, Giglio MG, Huttenhower C, Gevers D, Petrosino JF, Abubucker S, Badger JH, Chinwalla AT, Earl AM, FitzGerald MG, Fulton RS, Hallsworth-Pepin K, Lobos EA, Madupu R, Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren EJ, Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng Q, Aagaard KM, Abolude OO, *et al*: **A framework for human microbiome research.** *Nature* 2012, **486:**215–221.
20. Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, Schwager E, Knights D, Song SJ, Yassour M, Morgan XC, Kostic AD, Luo C, González A, McDonald D, Haberman Y, Walters T, Baker S, Rosh J, Stephens M, Heyman M, Markowitz J, Baldassano R, Griffiths A, Sylvester F, Mack D, Kim S, Crandall W, Hyams J, Huttenhower C, *et al*: **The treatment-naive microbiome in new-onset Crohn's disease.** *Cell Host Microbe* 2014, **15:**382–392.
21. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81:**559–575.
22. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R: **QIIME allows analysis of high-throughput community sequencing data.** *Nat Methods* 2010, **7:**335–336.
23. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P: **An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea.** *ISME J* 2012, **6:**610–618.
24. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R: **UniFrac: an effective distance metric for microbial community comparison.** *ISME J* 2011, **5:**169–172.
25. Meyer PE: **infotheo: Information-Theoretic Measures** [http://cran.r-project.org/web/packages/infotheo/index.html]
26. Press WH, Flannery BP, Teukolsky SA, Vetterling WT: **Statistical Description of Data.** In *Numerical Recipes in C: The Art of Scientific Computing.* 3rd edition. New York: Cambridge University Press; 1992:634..
27. Csardi G, Nepusz T: **The igraph software package for complex network research.** *InterJournal* 2006, **Complex Sy:**1695.
28. Knights D, Parfrey LW, Zaneveld J, Lozupone C, Knight R: **Human-associated microbial signatures: examining their predictive value.** *Cell Host Microbe* 2011, **10:**292–296.
29. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, Boucher G, Ripke S, Ellinghaus D, Burtt N, Fennell T, Kirby A, Latiano A, Goyette P, Green T, Halfvarson J, Haritunians T, Korn JM, Kuruvilla F, Lagacé C, Neale B, Lo KS, Schumm P, Törkvist L, Dubinsky MC, Brant SR, Silverberg MS, Duerr RH, Altshuler D, Gabriel S, *et al*: **Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease.** *Nat Genet* 2011, **43:**1066–1073.
30. Zuberi K, Franz M, Rodriguez H, Montojo J, Lopes CT, Bader GD, Morris Q: **GeneMANIA prediction server 2013 update.** *Nucleic Acids Res* 2013, **41:**W115–W122.
31. Migone TS, Zhang J, Luo X, Zhuang L, Chen C, Hu B, Hong JS, Perry JW, Chen SF, Zhou JXH, Cho YH, Ullrich S, Kanakaraj P, Carrell J, Boyd E, Olsen HS, Hu G, Pukac L, Liu D, Ni J, Kim S, Gentz R, Feng P, Moore PA, Ruben SM, Wei P: **TL1A is a TNF-like ligand for DR3 and TR6/DcR3 and functions as a T cell costimulator.** *Immunity* 2002, **16:**479–492.
32. Jin S, Chin J, Seeber S, Niewoehner J, Weiser B, Beaucamp N, Woods J, Murphy C, Fanning A, Shanahan F, Nally K, Kajekar R, Salas A, Planell N, Lozano J, Panes J, Parmar H, Demartino J, Narula S, Thomas-Karyat DA: **TL1A/TNFSF15 directly induces proinflammatory cytokines, including TNFα, from CD3 + CD161+ T cells to exacerbate gut inflammation.** *Mucosal Immunol* 2012, **6:**886–899.
33. Garrett WS, Gallini CA, Yatsunenko T, Michaud M, DuBois A, Delaney ML, Punit S, Karlsson M, Bry L, Glickman JN, Gordon JI, Onderdonk AB, Glimcher LH: **Enterobacteriaceae act in concert with the gut microbiota to induce spontaneous and maternally transmitted colitis.** *Cell Host Microbe* 2010, **8:**292–300.
34. Mylonaki M, Rayment NB, Rampton DS, Hudspith BN, Brostoff J: **Molecular characterization of rectal mucosa-associated bacterial flora in inflammatory bowel disease.** *Inflamm Bowel Dis* 2005, **11:**481–487.
35. Tyler AD, Milgrom R, Stempak JM, Xu W, Brumell JH, Muise AM, Sehgal R, Cohen Z, Koltun W, Shen B, Silverberg MS: **The NOD2insC polymorphism is associated with worse outcome following ileal pouch-anal anastomosis for ulcerative colitis.** *Gut* 2013, **62:**1433–1439.
36. Sehgal R, Berg A, Hegarty JP, Kelly AA, Lin Z, Poritz LS, Koltun WA: **NOD2/CARD15 mutations correlate with severe pouchitis after ileal pouch-anal anastomosis.** *Dis Colon Rectum* 2010, **53:**1487–1494.
37. Ben-Shachar S, Yanai H, Baram L, Elad H, Meirovithz E, Ofer A, Brazowski E, Tulchinsky H, Pasmanik-Chor M, Dotan I: **Gene expression profiles of ileal inflammatory bowel disease correlate with disease phenotype and advance understanding of its immunopathogenesis.** *Inflamm Bowel Dis* 2013, **19:**2509–2521.
38. Tyler AD, Milgrom R, Xu W, Stempak JM, Steinhart AH, McLeod RS, Greenberg GR, Cohen Z, Silverberg MS: **Antimicrobial antibodies are associated with a Crohn's disease-like phenotype after ileal pouch-anal anastomosis.** *Clin Gastroenterol Hepatol* 2012, **10:**507–12.e1.
39. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, Sinha R, Gilroy E, Gupta K, Baldassano R, Nessel L, Li H, Bushman FD, Lewis JD: **Linking long-term dietary patterns with Gut Microbial Enterotypes.** *Science* 2011, **334:**105–108.
40. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ,

Turnbaugh PJ: **Diet rapidly and reproducibly alters the human gut microbiome.** *Nature* 2013, **505:**559–563.
41. Eklund A: **beeswarm: The bee swarm plot, an alternative to stripchart** [http://cran.r-project.org/web/packages/beeswarm/index.html]
42. Bach J-F: **The effect of infections on susceptibility to autoimmune and allergic diseases.** *N Engl J Med* 2002, **347:**911–920.

doi:10.1186/s13073-014-0107-1
**Cite this article as:** Knights *et al.*: Complex host genetics influence the microbiome in inflammatory bowel disease. *Genome Medicine* 2014 **6**:107.