Mutation spectrum of *NOD2* reveals recessive inheritance as a main driver of Early Onset Crohn's Disease

Julie E. Horowitz[1], Neil Warner[2, 3], Jeffrey Staples[1], Eileen Crowley [2, 3], Ryan Murchie[2, 3], Cristopher Van Hout[1], Alejandra Klauer King[1], Karoline Fiedler[2, 3], Jeffrey G. Reid[1], John D. Overton[1], Alan R. Shuldiner[1], Aris Baras[1], Geisinger-Regeneron DiscovEHR Collaboration, Frederick E. Dewey[1], Anne Griffiths[2], Omri Gottesman[1], Aleixo M. Muise[2, 3, 4, *], Claudia Gonzaga-Jauregui[1, *]

[1] Regeneron Genetics Center, Regeneron Pharmaceuticals Inc., Tarrytown, NY, USA

[2] SickKids Inflammatory Bowel Disease Center, Hospital for Sick Children, Toronto, ON, Canada

[3] Cell Biology Program, Research Institute, Hospital for Sick Children, Toronto, ON, Canada.

[4] Departments of Pediatrics and Biochemistry, University of Toronto, Toronto, ON, Canada.

*Co-corresponding authors

Correspondence should be addressed to:

Aleixo Muise MD, Ph.D.
555 University Ave.
The Hospital for Sick Children
Toronto, ON, Canada, M5G 1X8
Email: aleixo.muise@utoronto.ca
Phone: 416-813-7735
Fax: 416-813-6531

Claudia Gonzaga-Jauregui, Ph.D.
Regeneron Genetics Center, Regeneron Pharmaceuticals Inc.
777 Saw Mill River Rd.
Tarrytown, NY, USA
Email: clau.gonzagajauregui@regeneron.com
Phone: 914-847-3451

Abstract:

Inflammatory bowel disease (IBD), clinically defined as Crohn's disease (CD), ulcerative colitis (UC), or IBD-unclassified, results in chronic inflammation of the gastrointestinal tract in genetically susceptible hosts. Pediatric onset IBD represents ≥25% of all IBD diagnoses and often presents with intestinal stricturing, perianal disease, and failed response to conventional treatments. *NOD2* was the first and is the most replicated locus associated with adult IBD, to date. To determine the role of *NOD2* and other genes in pediatric IBD, we performed whole-exome sequencing on a cohort of 1,183 patients with pediatric onset IBD (ages 0-18.5 years). We identified 92 probands who were homozygous or compound heterozygous for rare and low frequency *NOD2* variants accounting for approximately 8% of our cohort, suggesting a Mendelian recessive inheritance pattern of disease. Additionally, we investigated the contribution of recessive inheritance of *NOD2* alleles in adult IBD patients from the Regeneron Genetics Center (RGC)-Geisinger Health System DiscovEHR study, which links whole exome sequences to longitudinal electronic health records (EHRs) from 51,289 participants. We found that ~7% of cases in this adult IBD cohort, including ~10% of CD cases, can be attributed to recessive inheritance of *NOD2* variants, confirming the observations from our pediatric IBD cohort. Exploration of EHR data showed that 14% of these adult IBD patients obtained their initial IBD diagnosis before 18 years of age, consistent with early onset disease. Collectively, our findings show that recessive inheritance of rare and low frequency deleterious *NOD2* variants account for 7-10% of CD cases and implicate *NOD2* as a Mendelian disease gene for early onset Crohn's Disease.

Author Summary:

Pediatric onset inflammatory bowel disease (IBD) represents ≥25% of IBD diagnoses; yet the genetic architecture of early onset IBD remains largely uncharacterized. To investigate this, we performed whole-exome sequencing and rare variant analysis on a cohort of 1,183 pediatric onset IBD patients. We found that 8% of patients in our cohort were homozygous or compound heterozygous for rare or low frequency deleterious variants in the nucleotide binding and oligomerization domain containing 2 (*NOD2*) gene. Further investigation of whole-exome sequencing of a large clinical cohort of adult IBD patients uncovered recessive inheritance of rare and low frequency *NOD2* variants in 7% of cases and that the relative risk for *NOD2* variant homozygosity has likely been underestimated. While it has been reported that having >1 *NOD2* risk alleles is associated with increased susceptibility to Crohn's Disease (CD), our data formally demonstrate what has long been suspected: recessive inheritance of *NOD2* alleles is a mechanistic driver of early onset IBD, specifically CD, likely due to loss of NOD2 protein function. Our data suggest that a subset of IBD-CD patients with early disease onset is characterized by recessive inheritance of *NOD2* alleles, which has important implications for the screening, diagnosis, and treatment of IBD.

Introduction:

Inflammatory bowel disease (IBD) is a chronic inflammatory condition of the gastrointestinal (GI) tract that arises as part of an inappropriate response to commensal or pathogenic microbiota in a genetically susceptible individual (1-4). IBD encompasses Crohn's Disease (CD); ulcerative colitis (UC); and IBD unclassified (IBDU). The etiology of IBD is complex and has been attributed to defects in a number of cellular pathways including pathogen sensing, autophagy, maintenance of immune homeostasis, and intestinal barrier function, among other processes (3-18).

Great effort has been invested into defining the genetic factors that confer IBD susceptibility. To date, >200 unique loci have been associated with IBD through genome-wide association studies (GWAS), primarily in adult populations (19, 20). Nearly all the identified susceptibility loci exhibit low effect sizes (ORs ~1.0-1.5) individually (19), and collectively account for less than 20% of the heritable risk for IBD (19, 21). These observations support a complex disease model in which common variants of modest effect sizes interact with environmental factors including diet, smoking, and the intestinal microbiome (22, 23) to give rise to IBD susceptibility (4).

The earliest and most replicated genetic associations with IBD (24-26) correspond to a locus on chromosome 16 that encompasses the nucleotide-binding and oligomerization domain-containing 2 (*NOD2)* gene, with an average allelic odds ratio across multiple studies of *3.1* (19, 20). *NOD2* encodes an intracellular microbial sensor that recognizes muramyl dipeptide (MDP) motifs found on bacterial peptidoglycans (27, 28). Upon activation, NOD2 protein signals through the NF-κB family of proteins (29) to modulate transcription of genes encoding pro-inflammatory cytokines IL-8, TNF-α, and IL-1β (30-32), among others. Variation in *NOD2* accounts for approximately 20% of the genetic risk among CD cases, with three variants - p.R702W (ExAC MAF= 0.0227 across all populations), p.G908R (ExAC MAF= 0.0099), and

p.L1007fs (ExAC MAF= 0.0131) - accounting for over 80% of the disease-causing mutations in *NOD2* associated with adult CD (33), albeit not with UC; and particularly ileal versus colonic CD (34). These three "common" risk variants, typically observed in a heterozygous state, are predicted loss-of-function alleles that impair NF-κB activation in response to MDP ligands, *in vitro (28, 35-37)*.

With the assumption that genetic risk has a disproportionate effect over environmental risk in early onset disease, recent studies have focused on pediatric IBD cases (diagnosed <18y) (38). Pediatric IBD patients comprise 20-25% of all IBD cases and are typically more clinically severe than adult-onset patients, often exhibiting disease of the upper GI tract, small bowel inflammation, and perianal disease as well as failure to thrive and poor clinical response (4, 39). Results from GWAS conducted in this group of severely affected patients indicate that associated loci in early onset IBD significantly overlap with adult IBD loci, including both the *NOD2* locus and an additional 28 CD-specific loci previously implicated in adult-onset IBD (40-42). As the mechanism for these "common" IBD susceptibility loci in the pathogenesis of early onset IBD remains unclear (43), we performed whole-exome sequencing and rare variant analysis on a cohort of 1,183 pediatric onset IBD patients to elucidate the role of rare protein coding variation in IBD-associated genes, specifically *NOD2*, in this disease.

Results and Discussion:

We performed trio-based analysis of 492 complete trios using a proband-based analytical pipeline to identify all recessive (compound heterozygous and homozygous) and *de novo* variants of interest in the affected probands. In our initial analyses, we identified 10 families with recessive (compound heterozygous or homozygous), rare variants (2%≤MAF) in *NOD2*, all with a diagnosis of CD. We observed that some of the rare variants in these probands were inherited in *trans* from previously-reported CD risk alleles, mainly the p.G908R missense

variant. We identified two individuals who are compound heterozygous for the p.G908R risk allele in *trans* with a less common *NOD2* CD risk variant (p.N852S) in one case and a novel truncating indel (p.S506Vfs*73) in the second case (Table S1, Fam008 and Fam009). The observation of a CD-associated *NOD2* risk allele in *trans* from other rare or novel alleles led us to survey the rest of the probands, including singletons and those part of incomplete trios, for recessive inheritance, either in a homozygous or compound heterozygous manner, of *NOD2* variants, but expanding our allelic range to low-frequency variants (2%≤MAF≤5%). Through this approach we identified 108 probands with putative recessive *NOD2* variants. Visual inspection of sequence reads and orthogonal confirmation through Sanger sequencing excluded 13 probands with variants inherited in *cis* from an unaffected parent or heterozygous variants that were initially called as homozygous due to low coverage of the region and skewed allelic balance. Of note, we identified 5 probands carrying p.L1007fs and p.M863V risk variants, 4 of which were confirmed to occur in *cis* and were inherited from an unaffected parent. The remaining case with p.L1007fs and p.M863V was a singleton and thus phase could not be determined. These two variants segregate in *cis* within the same haplotype, as confirmed by segregation within the trios and as previously observed (44). Therefore, we excluded these 4 probands from our final count of recessively-inherited *NOD2* variants. Similarly, we identified 3 probands from 3 complete trios segregating the p.S431L and p.V793M reported risk variants in *cis* inherited from an unaffected carrier parent; these probands were also excluded. Three additional probands were excluded on the basis of a re-evaluation of the phenotype that excluded a clinical diagnosis of IBD.

Thus, we identified 92 probands with confirmed recessive *NOD2* variants within our pediatric onset IBD cohort. These include: 25 probands carrying homozygous variants, 41 probands with confirmed compound heterozygous variants, and an additional 26 singleton probands with putative compound heterozygous variants where phasing could not be performed (Table S1, Figure S1). The majority of the compound heterozygous individuals (65/67) carry a

known *NOD2* CD-risk allele in addition to either another known *NOD2* CD-risk allele or a novel *NOD2* variant, including some truncating loss-of-function variants supporting loss or impaired function of NOD2 in the pathophysiology of CD (6). In total, 92 of 1,183 (7.8%) of the probands in our pediatric onset IBD cohort conformed to a recessive, Mendelian inheritance mode for *NOD2* rare and low frequency (MAF≤5%) deleterious variants (Table 1, Figure 1, Table S1, and Table S4).

The 92 pediatric patients homozygous for *NOD2* mutations were predominantly male (71%) with a median age at diagnosis of 12.5 years (Table S1). At diagnosis, 83% displayed diagnostic features of Crohn's disease. 23% of the cohort displayed a constellation of extra-intestinal manifestations, mainly large joint arthritis, chronic recurrent multifocal osteomyelitis, recurrent fevers, erythema nodosum, and pyoderma gangrenosum. Only 6% of the cohort showed significant perianal disease (namely fistulae and abscesses, skin tags and fissures were not considered as perianal disease) (Table S1). Per the Montreal classification of IBD (45), 44% of the overall cohort of patients presented with ileal disease at diagnosis (L1). 25% presented with ileocolonic disease (L3) and 10% displayed features of colonic inflammation only (L2). Isolated upper disease was only present in 2% of the cohort (L4). We observed a progression of ileal disease in 21% (18.5% stricturing; 2.5% penetrating) with 21% requiring a resection. On review of the Crohn's disease patients only, 50% displayed L1 disease (terminal ileal +/- limited cecal disease), 32.9% L3 (ileocolonic) disease; 86.8% B1 (non-stricturing, non-penetrating) disease, 10.5% B2 (stricturing) disease (Table S1).

Given the substantial contribution of recessive *NOD2* variants to CD in our pediatric onset IBD cohort and the known contribution of *NOD2* to adult CD, we next investigated the contribution of *NOD2* recessivity in a large clinical population. For this, we examined a cohort of adult IBD patients from the RGC - Geisinger Health System (GHS) DiscovEHR study (46) . A key feature of the DiscovEHR study is the ability to link genomic sequence data to de-identified electronic health records (EHRs). Within this cohort, we identified 984 patients (of 51,289 total

sequenced DiscovEHR patient-participants) with a diagnosis of IBD, defined as having a problem list entry or an encounter diagnosis entered for two separate clinical encounters on separate calendar days for the ICD-9 codes 555* (Regional enteritis) or 556* (Ulcerative enterocolitis). For our analysis, we surveyed all instances of homozygous *NOD2* rare and low frequency variants (MAF≤5%); the same parameters applied to our pediatric IBD probands. Among patients with an IBD diagnosis, we identified 18 individuals who are either homozygous for the p.R702W risk allele (N=10) or homozygous for the p.L1007fs allele (N=8) (Table 2, Figure S2). We did not identify any p.G908R homozygous individuals with an IBD diagnosis in this cohort. Next, we looked for instances of putative compound heterozygosity among these adult IBD DiscovEHR patients. To investigate this, we searched for occurrences of two or more of the three most prevalent *NOD2* risk alleles (p.R702W, p.G908R, or p.L1007fs) in these individuals. We identified putative compound heterozygosity for the three main CD risk alleles, p.R702W/p.G908R (N=6), p.G908R/p.L1007fs (N=5), and p.R702W/p.L1007fs (N=11) (Table 2). We also observed instances of putative compound heterozygosity for each of the three main CD risk alleles along with either a rarer CD risk allele or a novel allele or two rare alleles in *trans* (N=24), parallel to the findings in our pediatric IBD cohort. Using familial relationships and pedigree reconstruction (47), we were able to confirm appropriate segregation for 32 of the 64 DiscovEHR recessive *NOD2* variant carriers with IBD, including *trans* inheritance in 13 putative compound heterozygotes (Figure S2). The other 32 were singleton cases where phase could not be confirmed. Overall, we identified 64 homozygous or putative compound heterozygous *NOD2* variant carriers in the DiscovEHR IBD cohort, accounting for 6.5% of patients with an IBD diagnosis in this clinical population (Figure 1, Table S4).

We were also able to evaluate longitudinal de-identified medical records for all patients within the DiscovEHR IBD cohort. According to their EHR data, 21 patients received diagnoses of both UC and CD. To clarify these diagnoses, we performed an evaluation of EHR information (which includes demographics, encounter and problem list diagnosis codes, procedure codes,

and medications) for all 64 homozygous or compound heterozygous *NOD2* patients with an IBD diagnosis. Through this review, 6 homozygotes exhibited a conflicting diagnosis of CD, of which 5 were resolved as CD and 1 could not be defined; 16 compound heterozygotes exhibited a conflicting diagnosis of CD of which 6 were resolved as CD and 10 were resolved as UC (Table S3). In total, we found that 17/18 (94.4%) of homozygous *NOD2* individuals and 33/46 (71.7%) compound heterozygous had a diagnosis of CD and that 9.9% of all CD cases in this cohort could be attributed to homozygous or compound heterozygous variants in *NOD2*. We next investigated age of disease onset using the first recorded date of an IBD diagnosis in the EHR. We identified 6 carriers of recessive *NOD2* variants (9.4% of our recessive *NOD2* patients with IBD) who were diagnosed with IBD prior to 18 years of age. We also identified an additional 11 carriers of recessive *NOD2* variants diagnosed with IBD prior to age 30 years, which is at or below the average age of IBD diagnosis (48) and is consistent with earlier disease onset (Table S3). Of note, our RGC-GHS DiscovEHR data extends to a median of 14 years (and maximum of 20 years) of electronically recorded medical information, concurrent with the adoption of the EHR by the Geisinger Health System. Since 72.4% of our cohort is currently over the age of 50 years, we cannot determine whether the age of onset for IBD occurred prior to the first electronically recorded date of an IBD diagnosis for many recessive *NOD2* patients; thus it is possible that other individuals with homozygous or compound heterozygous variants in *NOD2* might have had pediatric-onset disease that was not captured in the EHR.

Finally, given the recessive inheritance of *NOD2* variants observed in both our pediatric onset and adult IBD cohorts, we estimated the disease risk for the three main known CD risk alleles (p.R702W, p.G908R, and p.L1007fs) in our adult IBD case cohort and their effect sizes using additive, genotypic, and recessive genetic models. Under an additive model, we observed similar effect sizes for each of the 3 variants [OR=1.43 (1.20-1.71 95%CI, P-value $4.63X10^{-5}$) for p.R702W; OR=1.56 (1.18-2.06 95%CI, P-value $1.54X10^{-3}$) for p.G908R; and OR=1.84 (1.50-2.26 95%CI, P-value $1.69X10^{-9}$) for p.L1007fs], consistent with previously reported low to

moderate effect sizes for each allele by GWAS (19) and the most recent data available in the IBD Exomes Portal (49) (Table 3, Figure 2). However, for the two risk alleles with homozygous cases, in the genotypic model – which estimates distinct effect sizes for heterozygous and homozygous carriers – we observe substantially larger effects in homozygotes versus heterozygotes for the p.R702W variant (Het OR= 1.30 [1.06-1.58 95%CI], P-value $8.77X10^{-3}$, versus Hom OR= 4.02 [2.17-7.45 95% CI], P-value $6.86X10^{-6}$) and the p.L1007fs variant (Het OR= 1.63 [1.30-2.04 95%CI], P-value $1.67X10^{-5}$, versus Hom OR= 10.15 [4.75-21.69 95% CI], P-value $1.38X10^{-12}$). We also calculated the effect sizes using a recessive model for these two variants and the 22 compound heterozygotes carrying any combination of the 3 CD risk alleles. We found that recessive effect sizes for the p.R702W and p.L1007fs variants were similar to those observed under the homozygous genotypic model (OR=3.91 [2.11-7.24 95% CI], P-value $2.86X10^{-6}$, and OR=9.81 [4.59-20.94 95% CI], P-value $3.80X10^{-13}$, respectively) (Table 3, Figure 2). Further, under the recessive model we observed that the effect size for the compound heterozygotes was also significant (OR=4.35 [2.80-6.75 95% CI], P-value= $8.14x10^{-13}$), consistent with our previous observations (Table 3, Figure 2). Finally, we calculated the combined contribution of the 3 CD risk alleles under the different genetic models: additive (OR=1.64 [1.45-1.86 95%CI], P-value $4.58X10^{-15}$), genotypic (Het OR= 1.49 [1.28-1.73 95%CI], P-value $2.75X10^{-7}$, versus Hom OR= 5.24 [3.77-7.27 95% CI], P-value $4.31X10^{-22}$), and recessive (OR=4.81 [3.47-6.67 95% CI], P-value= $1.63x10^{-25}$) (Table 3, Figure 2). Collectively, these analyses show substantially larger effects for *NOD2* homozygotes and compound heterozygotes than heterozygotes and indicate that the genetic contribution of *NOD2* alleles, in a subset of IBD patients, is consistent with a recessive disease model.

These observations are in line with previous analyses and meta-analyses of CD cohorts where individuals carrying any one of the main three CD associated risk alleles (p.R702W, p.G908R, or p.L1007fs) have 2-4 fold increased risk for developing CD (36), whereas carriers of two or more of the same *NOD2* variants have a 15-40 fold increased risk for developing CD (33,

50, 51), exhibiting disease of the terminal ileum (34), and earlier diagnosis (by an average of 3 years) (33). Our observations support these studies but highlight a subset of IBD cases molecularly defined by recessive inheritance of *NOD2* alleles that exhibit markedly increased risk for CD with significantly earlier age of onset (mean age of onset among recessive *NOD2* carriers in the DiscovEHR IBD cohort: 43.4y; mean age of onset in the DiscovEHR IBD cohort: 51.5y; P-value: $4.0X10^{-4}$ by unpaired t test).

Further, while we observe a low effect size for single allele carriers, based on our allelic effect size calculations for each of the 3 main CD risk alleles in our DiscovEHR cohort (Table 3, Figure 2), we hypothesize that homozygous and compound heterozygous *NOD2* individuals included in large IBD GWAS cohorts have likely contributed to a large proportion of the relative risk calculations for IBD, specifically for CD, under additive models, and that homozygous effect sizes have been largely underappreciated or underreported. It is possible that stratification or conditional statistical analysis of these large and heterogeneous cohorts based on *NOD2* genotypes may increase power to detect other loci that contribute to IBD.

While our observations strongly support recessive inheritance of *NOD2* variants as a driver of early onset Crohn's disease, we observed incomplete penetrance, as evidenced by homozygous or compound heterozygous *NOD2* variant carriers that do not have a clinical presentation of IBD (52-54). Penetrance and expressivity are two major genetic concepts that play into the onset of the phenotype and the clinical presentation of monogenic diseases (55). In the case of IBD, penetrance is known to be incomplete and clinical presentation is extremely variable. Further, the contribution of additional environmental triggers that may enhance disease onset and/or severity in an already genetically-compromised individual should not be underestimated, especially considering that the loss of epithelial barrier function occurring during IBD allows for host exposure to up to 10^{14} gut microbiota (56, 57). Even in cases of monogenic IBD, such as IL-10 receptor deficiency (58-60), intestinal flora are required for disease presentation in murine disease models (61-63). Furthermore, variation in genes

involved in NOD2-dependent signaling pathways, including *XIAP* (64-66) and *TRIM22* (67), result in Mendelian forms of IBD. For *XIAP*, and most likely *TRIM22,* viral triggers are required for disease onset and progression, and *XIAP* mutations have variable penetrance, with only a small percentage of XIAP-deficiency patients developing CD (age of onset between 3 months and 40 years (52)). As *NOD2*-deficient hosts are more susceptible to the pathogenic effects of a changing intestinal microenvironment (68), the contribution of either discrete or continuous gene-environment exposures may further explain heterogeneity in onset and presentation of disease for genetically-sensitized recessive *NOD2* carriers.

Given the wide variability in clinical presentation of IBD (54), we cannot exclude the possibility that recessive *NOD2* carriers exhibit subclinical phenotypes not formally diagnosed as IBD or that they may eventually develop IBD. It is additionally possible that recessive *NOD2* carriers in the DiscovEHR cohort have a diagnosis of IBD that has not been captured in the EHR. Future investigation into the medical histories of recessive *NOD2* carriers may shed light on this variable expressivity or incomplete capture of medical information. We also cannot exclude the possibility that recessive *NOD2* carriers possess additional genes or alleles that either contribute to disease onset and severity or, alternatively, provide protection or reduced expressivity of the phenotype. Identification of these genetic modifiers warrants future investigation both to unveil additional IBD-risk associated loci for early onset UC and CD cases and to identify protective genes and alleles that can be used to derive therapeutic avenues for IBD treatment and management.

In summary, in a cohort of 1,183 pediatric and early onset IBD patients, we report recessive inheritance of rare and low frequency variants in *NOD2* accounting for about 8% of probands. We assessed the contribution of *NOD2* recessive inheritance in a broader, heterogeneous cohort of adult IBD patients, similar to those recruited for GWAS, and found that recessive inheritance of variants in *NOD2* account for 6.5% of these IBD patients, including 9.9% of CD cases. Thus, recessive inheritance of rare and low frequency *NOD2* variants

explain a substantial proportion of CD cases in a pediatric cohort and a large clinical population, with significantly earlier age of disease onset. Consistently, both pediatric and adult CD exhibit a broad spectrum of clinical presentation, suggesting a shared etiology across age groups, at least in the subgroup defined by recessive *NOD2*-driven CD. Our findings indicate that deleterious *NOD2* variants should be considered as strong predictors of IBD-CD onset and implicate *NOD2* as a Mendelian disease gene for early onset IBD, specifically for a molecularly defined subset of Crohn's disease patients.

Materials and Methods:

We performed whole exome sequencing on a cohort of 1,183 probands with pediatric onset IBD (ages 0-18.5 years), including their affected and unaffected parents and siblings, where available (total samples = 2,704). All individuals were consented for genetic studies under an IRB-approved protocol by the Toronto Hospital for Sick Children. Sample preparation, whole exome sequencing, and sequence data production were performed at the Regeneron Genetics Center (RGC) as previously described (46). In brief, 1ug of high-quality genomic DNA for exome capture using the NimbleGen VCRome 2.1 design. Captured libraries were sequenced on the Illumina HiSeq 2500 platform with v4 chemistry using paired-end 75 bp reads. Exome sequencing was performed such that >85% of the bases were covered at 20x or greater. Raw sequence reads were mapped and aligned to the GRCh37/hg19 human genome reference assembly, and called variants were annotated and analyzed using an RGC developed pipeline. Briefly, variants were filtered based on their observed minor allele frequencies at a <2% cutoff using the internal RGC database and other population control databases (ExAC (69) and NHLBI's ESP (70)) to filter out common polymorphisms and high frequency, likely benign variants in consideration of disease prevalence.

335 **Conflict of Interest Disclosure**

336 J.E.H. is a postdoctoral fellow at the Regeneron Genetics Center, Regeneron
337 Pharmaceuticals, Inc. J.S., C.V.H., A.K.K., J.G.R., J.D.O., A.R.S., A.B., F.E.D, O.G., and C.G.J
338 are full-time employees of the Regeneron Genetics Center, Regeneron Pharmaceuticals, Inc.
339 and receive stock options as part of compensation. N.W., R.M., K.F., A.G., and A.M.M have no
340 conflicts to disclose.

341

342 **Acknowledgements**

343 We are thankful to the patients and families who participated in this study. A.M.M. is funded by a
344 CIHR – Operating Grant (MOP119457) and the Leona M. and Harry B. Helmsley Charitable
345 Trust to study VEOIBD.

346

Figures and Tables

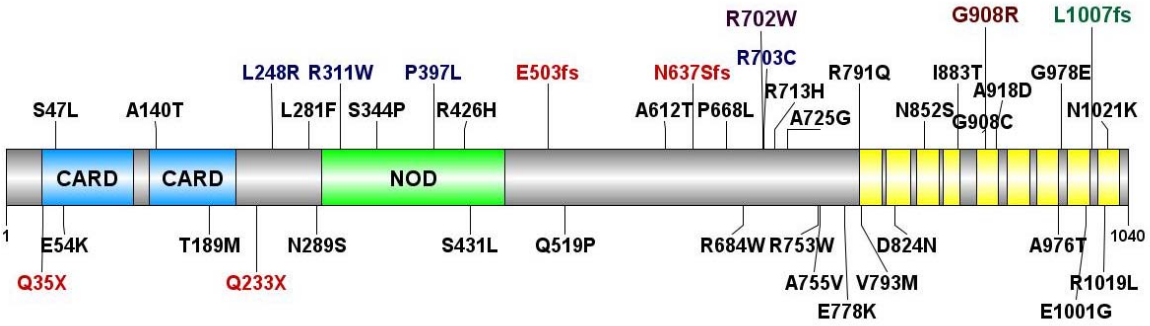


Figure 1. Mutation spectrum of *NOD2* in inflammatory bowel disease (IBD) patients. *NOD2* variation identified in patients with pediatric early onset IBD (upper) and adult IBD cohort from the RGC-GHS DiscovEHR collaboration (lower). Variants in blue were observed in both cohorts; variants in red are predicted loss-of-function that result in nonsense mediated decay. The three “common” low-frequency Crohn’s Disease risk variants are highlighted: R702W (purple), G908R (brown), and L1007fs (green). Also depicted are the NOD2 protein structural domains: two caspase activation and recruitment domains (CARD), a nucleotide binding and oligomerization (NOD) domain, and leucine rich repeat domains in yellow.

Table 1. Mutation spectrum of recessive *NOD2* variants in an EO-IBD cohort.

NOD2 Variant (p.)	# EO-IBD Probands	Mean Age (Range)	% CD Dx	Tissue Involvement		
				Colon	Ileum	Perianal
Compound Heterozygous						
Rare/Rare	4T; 1S	12.5 (9.0-14.6)	88.9%	80.0%	60.0%	40.0%
Common/Rare	1Q; 9T; 6D; 14S	12.6 (5.5-16.5)	93.1%	57.1%	82.1%	25.0%
Common/Common	17T; 5D; 10S	11.8 (2.1-18.5)	90.6%	56.3%	90.6%	25.0%
Homozygous						
Rare	1D; 2S	9.6 (4.2-15.1)	66.7%	33.3%	33.3%	33.3%
p.R702W	1Q; 2T; 3D; 1S	11.7 (5.8-13.8)	100%	85.7%	57.1%	42.9%
p.G908R	2T; 1D; 2S	10.6 (7.7-13.7	80.0%	40.0%	40.0%	0.0%
p.L1007fs	4T; 2D; 4S	12.1 (5.9-13.4)	100%	60.0%	90.0%	20.0%

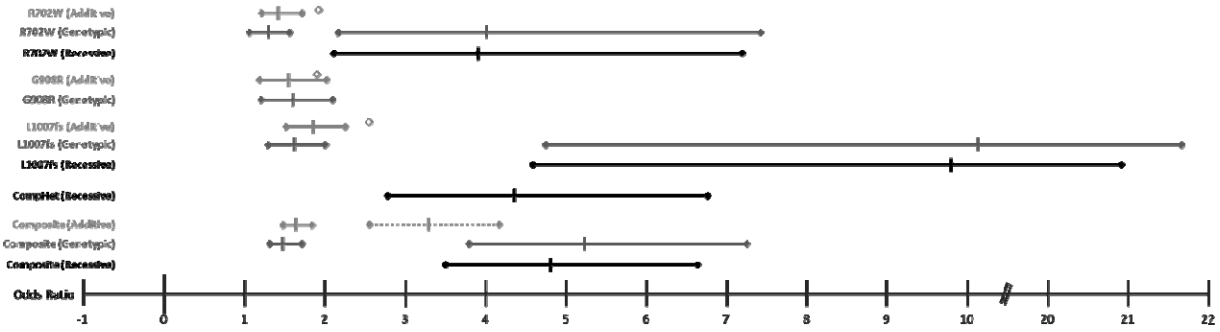
Common NOD2 variants refer to the three main low-frequency Crohn's Disease risk variants p.R702W, p.G908R, and p.L1007fs; Rare NOD2 variants refer to other low-frequency variants (MAF≤5%). Abbreviations: Q, quartet; T, trio; D, duo; S, singleton; Dx, diagnosis

Table 2. Mutation spectrum of recessive *NOD2* variants in the RGC-GHS DiscovEHR adult IBD cohort.

NOD2 Variants	# IBD patients	Mean Age (Range)	# CD Dx	% CD Dx	# Male	% Male	# Perianal	% Perianal
Homozygous								
p.R702W	10	47.3 (16.0-76.3)	10	100%	5	50.0%	1	10.0%
p.G908R	0	-	-	-	-	-	-	-
p.L1007fs	8	40.25 (11.0-75.0)	8	100%	4	50.0%	3	37.5%
Compound Heterozygous								
Common/Common								
p.R702W/p.G908R	6	48.5 (11.4-54.2)	3	50.0%	5	83.3%	1	16.7%
p.R702W/p.L1007fs	11	38.5 (21.5-69.2)	8	72.7%	3	27.3%	2	18.2%
p.G908R/p.L1007fs	5	35.2 (20.0-52.4)	4	80.0%	2	40.0%	0	0.0%
Common/Rare	16	40.3 (10.8-66.1)	8	50.0%	8	50.0%	6	37.5%
Rare/Rare	8	60.9 (30.8-78.7)	4	50.0%	3	37.5%	1	12.5%

Common NOD2 variants refer to the three main low-frequency Crohn's Disease risk variants p.R702W, p.G908R, and p.L1007fs; Rare NOD2 variants refer to other low-frequency variants (MAF≤5%). Abbreviations: Dx, diagnosis.

Figure 2. Graphical representation of Odds Ratio (OR) point estimates and 95% confidence intervals (CI) for the three main CD risk alleles (p.R702W, p.G908R, p.L1007fs) under additive, genotypic, and recessive genetic models (corresponding to values in Table 3). The dotted line in the Composite panel depicts the calculated CI with corresponding calculated OR for 2 alleles under an additive genetic model; of note the point estimate (2xOR) is outside of the 95% CI for the Composite genotypic homozygous and recessive models. Diamonds correspond to estimated OR values for these same variants in the IBD Exomes Browser(49); no confidence intervals are provided.



382 **Table 3. OR calculations for 3 NOD2 CD risk alleles (p.R702W, p.G908R, p.L1007fs),**
383 **composite, and compound heterozygous combinations in the DiscovEHR cohort.** There
384 were no homozygotes for the p.G908R variant affected with IBD in our cohort; therefore, no
385 genotypic homozygous and recessive ORs could be calculated. The 'Composite *NOD2*'
386 calculations account for all alleles and genotypes for the 3 CD risk variants in the different
387 genetic models.

NOD2 Variant	DiscovEHR MAF	DiscovEHR Controls (N=50,305)	DiscovEHR IBD Cases (N=984)	Additive Model OR [95% CI] (*P*-value)	Genotypic Model (Heterozygous) OR [95% CI] (*P*-value)	Genotypic Model (Homozygous) OR [95% CI] (*P*-value)	Recessive Model OR [95% CI] (*P*-value)	ExAC MAF	IBD Exomes OR (*P*-value)
p.R702W	0.050	Het=4,843; Hom=156	Het=116; Hom=10	1.43 [1.20-1.71] ($4.63x10^{-5}$)	1.30 [1.06-1.58] (0.008765)	4.02 [2.17-7.45] ($6.86x10^{-6}$)	3.91 [2.11-7.24] ($2.86x10^{-6}$)	0.035	1.92 ($<1x10^{-16}$)
p.G908R	0.017	Het=1,735; Hom=16	Het=52; Hom=0	1.56 [1.18-2.06] (0.001544)	1.61 [1.21-2.13] (0.00087)	NA	NA	0.012	1.91 ($<1x10^{-16}$)
p.L1007fs	0.029	Het=2,894; Hom=50	Het=86; Hom=8	1.84 [1.50-2.26] ($1.69x10^{-9}$)	1.63 [1.30-2.04] ($1.67x10^{-5}$)	10.15 [4.75-21.69] ($1.38x10^{-12}$)	9.80 [4.59-20.94] ($3.80x10^{-13}$)	0.018	2.57 ($<1x10^{-16}$)
Compound Heterozygous	.	N=285	N=22	.	.	.	4.35 [2.80-6.75] ($8.14x10^{-13}$)	.	.
Composite *NOD2*	.	Het=8955; Rec=450	Het=232; Rec=41	1.64 [1.45-1.86] ($4.58x10^{-15}$) 3.29 [2.56-4.23] (2 alleles predicted)	1.49 [1.28-1.73] ($2.75x10^{-7}$)	5.24 [3.77-7.27] ($4.31x10^{-22}$)	4.81 [3.47-6.67] ($1.63x10^{-25}$)	.	.

388

389

References

1. Abraham C, Cho JH. Inflammatory bowel disease. N Engl J Med. 2009;361(21):2066-78.
2. Van Limbergen J, Wilson DC, Satsangi J. The genetics of Crohn's disease. Annu Rev Genomics Hum Genet. 2009;10:89-116.
3. Cho JH. The genetics and immunopathogenesis of inflammatory bowel disease. Nature reviews Immunology. 2008;8(6):458-66.
4. Kaser A, Zeissig S, Blumberg RS. Inflammatory bowel disease. Annu Rev Immunol. 2010;28:573-621.
5. Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet. 2009;41(6):703-7.
6. Cho JH, Abraham C. Inflammatory bowel disease genetics: Nod2. Annu Rev Med. 2007;58:401-16.
7. Abraham C, Cho JH. Functional consequences of NOD2 (CARD15) mutations. Inflamm Bowel Dis. 2006;12(7):641-50.
8. Chu H, Khosravi A, Kusumawardhani IP, Kwon AH, Vasconcelos AC, Cunha LD, et al. Gene-microbiota interactions contribute to the pathogenesis of inflammatory bowel disease. Science. 2016;352(6289):1116-20.
9. Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, Huse K, et al. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nat Genet. 2007;39(2):207-11.
10. Parkes M, Barrett JC, Prescott NJ, Tremelling M, Anderson CA, Fisher SA, et al. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat Genet. 2007;39(7):830-2.
11. Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A, et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet. 2007;39(5):596-604.
12. Abraham C, Cho JH. IL-23 and autoimmunity: new insights into the pathogenesis of inflammatory bowel disease. Annu Rev Med. 2009;60:97-110.
13. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science. 2006;314(5804):1461-3.
14. Cattin AL, Le Beyec J, Barreau F, Saint-Just S, Houllier A, Gonzalez FJ, et al. Hepatocyte nuclear factor 4alpha, a key factor for homeostasis, cell architecture, and barrier function of the adult intestinal epithelium. Mol Cell Biol. 2009;29(23):6294-308.
15. Khor B, Gardet A, Xavier RJ. Genetics and pathogenesis of inflammatory bowel disease. Nature. 2011;474(7351):307-17.
16. Muise AM, Walters TD, Glowacka WK, Griffiths AM, Ngan BY, Lan H, et al. Polymorphisms in E-cadherin (CDH1) result in a mis-localised cytoplasmic protein that is associated with Crohn's disease. Gut. 2009;58(8):1121-7.
17. Scharl M, Paul G, Weber A, Jung BC, Docherty MJ, Hausmann M, et al. Protection of epithelial barrier function by the Crohn's disease associated gene protein tyrosine phosphatase n2. Gastroenterology. 2009;137(6):2030-40 e5.
18. Kaser A, Lee AH, Franke A, Glickman JN, Zeissig S, Tilg H, et al. XBP1 links ER stress to intestinal inflammation and confers genetic risk for human inflammatory bowel disease. Cell. 2008;134(5):743-56.
19. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491(7422):119-24.

20. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet. 2015;47(9):979-86.
21. McGovern DP, Kugathasan S, Cho JH. Genetics of Inflammatory Bowel Diseases. Gastroenterology. 2015;149(5):1163-76 e2.
22. Flint HJ, Scott KP, Louis P, Duncan SH. The role of the gut microbiota in nutrition and health. Nat Rev Gastroenterol Hepatol. 2012;9(10):577-89.
23. Kostic AD, Xavier RJ, Gevers D. The microbiome in inflammatory bowel disease: current status and the future ahead. Gastroenterology. 2014;146(6):1489-99.
24. Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature. 2001;411(6837):599-603.
25. Hugot JP, Laurent-Puig P, Gower-Rousseau C, Olson JM, Lee JC, Beaugerie L, et al. Mapping of a susceptibility locus for Crohn's disease on chromosome 16. Nature. 1996;379(6568):821-3.
26. Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, et al. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature. 2001;411(6837):603-6.
27. Girardin SE, Boneca IG, Viala J, Chamaillard M, Labigne A, Thomas G, et al. Nod2 is a general sensor of peptidoglycan through muramyl dipeptide (MDP) detection. J Biol Chem. 2003;278(11):8869-72.
28. Inohara N, Ogura Y, Fontalba A, Gutierrez O, Pons F, Crespo J, et al. Host recognition of bacterial muramyl dipeptide mediated through NOD2. Implications for Crohn's disease. J Biol Chem. 2003;278(8):5509-12.
29. Van Limbergen J, Radford-Smith G, Satsangi J. Advances in IBD genetics. Nat Rev Gastroenterol Hepatol. 2014;11(6):372-85.
30. Inohara, Chamaillard, McDonald C, Nunez G. NOD-LRR proteins: role in host-microbial interactions and inflammatory disease. Annu Rev Biochem. 2005;74:355-83.
31. Strober W, Murray PJ, Kitani A, Watanabe T. Signalling pathways and molecular interactions of NOD1 and NOD2. Nat Rev Immunol. 2006;6(1):9-20.
32. Watanabe T, Kitani A, Strober W. NOD2 regulation of Toll-like receptor responses and the pathogenesis of Crohn's disease. Gut. 2005;54(11):1515-8.
33. Lesage S, Zouali H, Cezard JP, Colombel JF, Belaiche J, Almer S, et al. CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. Am J Hum Genet. 2002;70(4):845-57.
34. Cleynen I, Boucher G, Jostins L, Schumm LP, Zeissig S, Ahmad T, et al. Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. Lancet. 2016;387(10014):156-67.
35. Bonen DK, Ogura Y, Nicolae DL, Inohara N, Saab L, Tanabe T, et al. Crohn's disease-associated NOD2 variants share a signaling defect in response to lipopolysaccharide and peptidoglycan. Gastroenterology. 2003;124(1):140-6.
36. Chamaillard M, Philpott D, Girardin SE, Zouali H, Lesage S, Chareyre F, et al. Gene-environment interaction modulated by allelic heterogeneity in inflammatory diseases. Proc Natl Acad Sci U S A. 2003;100(6):3455-60.
37. van Heel DA, Ghosh S, Butler M, Hunt KA, Lundberg AM, Ahmad T, et al. Muramyl dipeptide and toll-like receptor sensitivity in NOD2-associated Crohn's disease. Lancet. 2005;365(9473):1794-6.
38. Moran CJ, Klein C, Muise AM, Snapper SB. Very early-onset inflammatory bowel disease: gaining insight through focused discovery. Inflamm Bowel Dis. 2015;21(5):1166-75.
39. Rosen MJ, Dhawan A, Saeed SA. Inflammatory Bowel Disease in Children and Adolescents. JAMA Pediatr. 2015;169(11):1053-60.

40. Cutler DJ, Zwick ME, Okou DT, Prahalad S, Walters T, Guthery SL, et al. Dissecting Allele Architecture of Early Onset IBD Using High-Density Genotyping. PLoS One. 2015;10(6):e0128074.
41. Imielinski M, Baldassano RN, Griffiths A, Russell RK, Annese V, Dubinsky M, et al. Common variants at five new loci associated with early-onset inflammatory bowel disease. Nat Genet. 2009;41(12):1335-40.
42. Kugathasan S, Baldassano RN, Bradfield JP, Sleiman PM, Imielinski M, Guthery SL, et al. Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. Nature genetics. 2008;40(10):1211-5.
43. Paul T, Birnbaum A, Pal DK, Pittman N, Ceballos C, LeLeiko NS, et al. Distinct phenotype of early childhood inflammatory bowel disease. J Clin Gastroenterol. 2006;40(7):583-6.
44. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. 2011;43(11):1066-73.
45. Silverberg MS, Satsangi J, Ahmad T, Arnott ID, Bernstein CN, Brant SR, et al. Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: report of a Working Party of the 2005 Montreal World Congress of Gastroenterology. Can J Gastroenterol. 2005;19 Suppl A:5A-36A.
46. Dewey FE, Murray MF, Overton JD, Habegger L, Leader JB, Fetterolf SN, et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. Science. 2016;354(6319).
47. Staples J, Qiao D, Cho MH, Silverman EK, Nickerson DA, Below JE. PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. Am J Hum Genet. 2014;95(5):553-64.
48. Cosnes J, Gower-Rousseau C, Seksik P, Cortot A. Epidemiology and natural history of inflammatory bowel diseases. Gastroenterology. 2011;140(6):1785-94.
49. Portal IE. Cambridge, MA [Available from: http://ibd.broadinstitute.org/.
50. Economou M, Trikalinos TA, Loizou KT, Tsianos EV, Ioannidis JP. Differential effects of NOD2 variants on Crohn's disease risk and phenotype in diverse populations: a metaanalysis. Am J Gastroenterol. 2004;99(12):2393-404.
51. Hugot JP, Zaccaria I, Cavanaugh J, Yang H, Vermeire S, Lappalainen M, et al. Prevalence of CARD15/NOD2 mutations in Caucasian healthy people. Am J Gastroenterol. 2007;102(6):1259-67.
52. Uhlig HH, Schwerd T, Koletzko S, Shah N, Kammermeier J, Elkadri A, et al. The diagnostic approach to monogenic very early onset inflammatory bowel disease. Gastroenterology. 2014;147(5):990-1007 e3.
53. Kammermeier J, Dziubak R, Pescarin M, Drury S, Godwin H, Reeve K, et al. Phenotypic and genotypic characterisation of inflammatory bowel disease presenting before the age of two years. J Crohns Colitis. 2016.
54. Ray A, Dittel BN. Interrelatedness between dysbiosis in the gut microbiota due to immunodeficiency and disease penetrance of colitis. Immunology. 2015;146(3):359-68.
55. Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum Genet. 2013;132(10):1077-130.
56. Mankertz J, Schulzke JD. Altered permeability in inflammatory bowel disease: pathophysiology and clinical implications. Curr Opin Gastroenterol. 2007;23(4):379-83.
57. Guinane CM, Cotter PD. Role of the gut microbiota in health and chronic gastrointestinal disease: understanding a hidden metabolic organ. Therap Adv Gastroenterol. 2013;6(4):295-308.

541 58. Glocker EO, Kotlarz D, Boztug K, Gertz EM, Schaffer AA, Noyan F, et al. Inflammatory
542 bowel disease and mutations affecting the interleukin-10 receptor. N Engl J Med.
543 2009;361(21):2033-45.
544 59. Kotlarz D, Beier R, Murugan D, Diestelhorst J, Jensen O, Boztug K, et al. Loss of
545 interleukin-10 signaling and infantile inflammatory bowel disease: implications for diagnosis and
546 therapy. Gastroenterology. 2012;143(2):347-55.
547 60. Muise AM, Snapper SB, Kugathasan S. The age of gene discovery in very early onset
548 inflammatory bowel disease. Gastroenterology. 2012;143(2):285-8.
549 61. Kuhn R, Lohler J, Rennick D, Rajewsky K, Muller W. Interleukin-10-deficient mice
550 develop chronic enterocolitis. Cell. 1993;75(2):263-74.
551 62. Sellon RK, Tonkonogy S, Schultz M, Dieleman LA, Grenther W, Balish E, et al. Resident
552 enteric bacteria are necessary for development of spontaneous colitis and immune system
553 activation in interleukin-10-deficient mice. Infect Immun. 1998;66(11):5224-31.
554 63. Yang I, Eibach D, Kops F, Brenneke B, Woltemate S, Schulze J, et al. Intestinal
555 microbiota composition of interleukin-10 deficient C57BL/6J mice and susceptibility to
556 Helicobacter hepaticus-induced colitis. PLoS One. 2013;8(8):e70783.
557 64. Abbott DW, Yang Y, Hutti JE, Madhavarapu S, Kelliher MA, Cantley LC. Coordinated
558 regulation of Toll-like receptor and NOD2 signaling by K63-linked polyubiquitin chains. Mol Cell
559 Biol. 2007;27(17):6012-25.
560 65. Damgaard RB, Fiil BK, Speckmann C, Yabal M, zur Stadt U, Bekker-Jensen S, et al.
561 Disease-causing mutations in the XIAP BIR2 domain impair NOD2-dependent immune
562 signalling. EMBO Mol Med. 2013;5(8):1278-95.
563 66. Krieg A, Correa RG, Garrison JB, Le Negrate G, Welsh K, Huang Z, et al. XIAP
564 mediates NOD signaling via interaction with RIP2. Proc Natl Acad Sci U S A.
565 2009;106(34):14524-9.
566 67. Li Q, Lee CH, Peters LA, Mastropaolo LA, Thoeni C, Elkadri A, et al. Variants in TRIM22
567 That Affect NOD2 Signaling Are Associated With Very-Early-Onset Inflammatory Bowel
568 Disease. Gastroenterology. 2016;150(5):1196-207.
569 68. Ramanan D, Bowcutt R, Lee SC, Tang MS, Kurtz ZD, Ding Y, et al. Helminth infection
570 promotes colonization resistance via type 2 immunity. Science. 2016;352(6285):608-12.
571 69. Lek M. Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv [Internet].
572 2015. Available from: http://biorxiv.org/content/early/2015/10/30/030338.
573 70. Exome Variant Server [Internet]. [cited July, 2016]. Available from:
574 http://evs.gs.washington.edu/EVS/.
575