Intestinal cancer development in response to oral infection with high-fat diet induced type 2 diabetes in collaborative cross mice under different host genetic background effects

Iqbal M. Lone
Tel-Aviv University
Asal Milhem
Tel-Aviv University
Nadav Ben Nun
Tel-Aviv University
Fuad A. Iraqi (fuadi@tauex.tau.ac.il)
Tel-Aviv University

Research Article

Keywords: Type 2 diabetes (T2D), High-fat diet (HFD), Collaborative cross (CC) mouse model, Polyp count, Machine learning

Posted Date: June 21st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1749983/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Mammalian Genome on February 9th, 2023. See the published version at https://doi.org/10.1007/s00335-023-09979-y.

Abstract

Background: Type 2 diabetes (T2D) is a metabolic disease with an imbalance in blood glucose concentration. There are significant studies currently showing association between T2D and intestinal cancer developments. High fat diet (HFD) plays part in the disease development of T2D, intestinal cancer, and infectious diseases through many biological mechanisms, including, but not limited to inflammation. Understanding the systems genetics of the multimorbidity of these diseases will provide an important knowledge and platform for dissecting the complexity of these diseases. Furthermore, in this study we used some machine learning (ML) models to explore more aspects of diabetes mellitus.

Aims: The ultimate aim of this project is to study the genetic factors, which underline T2D development, associated with intestinal cancer in response to a HFD consumptions and oral co-infection, jointly or separately, on the same host genetic background.

Materials & Methods: A cohort of 307 mice of eight different CC mouse lines in the four experimental groups was assessed. The mice were maintained on either HFD or Chow diet (CHD) for 12 weeks period, while half of each dietary group was either co-infected with oral bacteria or un-infected. Host response to a glucose load and clearance was assessed using intraperitoneal glucose tolerance test (IPGTT) at two time points (weeks 6 and 12) during the experiment period and, subsequently was translated to area under curve (AUC) values. At week 5 of the experiment, mice of group two and four were co-infected with *Porphyromonas gingivalis* (Pg) *and Fusobacterium nucleatum* (Fn) strains, three times a week, while keeping the other uninfected mice as a control group. At week 12, mice were sacrificed, small intestines and colon were extracted and subsequently the Polyp counts assessed, as well the intestine lengths and size measured.

Results & conclusions: Our results have shown that there is a significant variation in polyp's number in different CC lines, with a spectrum between 2.5 and 12.8 total polyps on average. There was a significant correlation between area under curve (AUC) and intestine measurements, including polyp counts, length and size. In addition, our results have shown a significant sex-effect on polyp development and glucose tolerance ability with males more susceptible to HFD than females by showing higher AUC in the glucose tolerance test. The ML results showed that classification with random forest could reach the highest accuracy when all the attributes were used.

These results provide an excellent platform for proceeding towards understanding the nature of the genes involved in resistance and rate of development of intestinal cancer and T2D induced by HFD and oral co-infection. Once obtained, such data can be used to predict individual risk for developing these diseases and to establish the genetically based strategy for their prevention and treatment.

Introduction

Type 2 diabetes (T2D) is a complex chronic disease characterized by impaired glucose tolerance by developing high insulin resistance in the targeted organs (Lone and Iraqi 2022; Kasuga 2006). T2D is

affected by genetic and environmental factors in particular by diet. Previous studies have found correlations between the consumption of high-calorie western diet concurrence with a sedentary lifestyle and the presence of T2D (Kolb and Martin 2017). T2D is considered as a systemic chronic disease and affects many host organs, which are involved in the disease development. This mainly includes pancreas (β cells and α cells), liver, skeletal muscles, kidneys, brain, small intestines, and adipose tissues (Defronzo 2009). It was found that T2D may affect changes in the colon microbiome, immune dysregulation, and inflammation and therefore emerged as pathophysiological factors prominently (Blasco-Baque et al. 2017).

Nearly 20% of malignancies in humans can be linked to infectious agents. The chronic infection was also found to enhance cancer progression through activating tumor-promoting signaling pathways (NF-kB, STAT3). This may increase supplementing the production of growth factors, antiapoptotic proteins and cytokines that foster cancer growth, dissemination and strengthen resistance to therapy (Atanasova et al. 2014; Szczepanski et al. 2009). Previous studies have shown within this context, that *Fusobacterium nucleatum* (Fn) and *Porphyromonas gingivalis* (Pg) both are oral bacteria, trigger Toll-like receptor (TLR) signaling in pre-cancerous/cancerous epithelium resulting in overexpression of epithelial-derived IL-6 (Rakoff-Nahoum et al. 2009; Whitmore et al. 2014).

Colorectal cancer (CRC) is the third deadliest and fourth most commonly diagnosed cancer in the world and has shown a steady rise in its incidence (American Cancer Society 2020). Based on the year 2020 records in United States, 104,610 CRC cases were reported and expected to cause about 53,200 deaths. About 1 in 23 (4.4%) men have a lifetime risk of developing CRC, and 1 in 25 women (4.1%), showing the risk is only slightly lower for women than for men (American Cancer Society 2014).

Obesity and metabolic status were shown to be linked through epidemiological and molecular evidence to inflammation, and an increased risk of many cancers, as well as periodontitis (Tan Chen 2016). Unhealthy and sedentary life-styles lacking physical activity, accompanied with high fat and caloric density as well as lacking in fruits and vegetables, high alcoholic consumption and long-term smoking are considered as leading risk factors for CRC development (Campbell 2010).

Previous epidemiological studies showed that individuals suffering from T2D might have a higher risk of developing CRC, when compared with their non-diabetic counterparts (Cavicchia et al. 2013; Peeters 2015; Tabak et al. 2009). It was proposed that T2D-CRC connection was, hypothetically explained by elevation of insulin resistance and hyperinsulinaemia, which occurs in early stages of T2D (Giovannucci 1995). Colorectal carcinogenesis is indirectly promoted by insulin through the increase of bioavailable insulin-like growth factor 1, which may enhance cell proliferation as well as inhibit apoptosis. Insulin is also an important growth factor of colonic epithelial cells and is a mitogen of tumor cell growth in vitro (Gallagher and LeRoith 2010; Renehan et al. 2015).

High fat diet (HFD) consumption is a common, prominent risk factor for T2D and intestinal cancer developments speculated to play its part through biological mechanisms including inflammation (Iyengar et al. 2015; Jiang et al. 2016; Muluke et al. 2016). Inflammatory cytokines are, particularly correlated with

elevated levels of saturated fatty acids (SFA). Previous studies from our research team have shown that increased body weight, waist circumference, body mass index (BMI), fasting glucose levels, and impaired glucose tolerance were induced by maintaining the experimental animals on HFD (Atamni et al. 2016a). As shown in an experimental study, SFAs unlike unsaturated fatty acids, have inflammatory potential. This accelerates alveolar bone loss in PD in obese mice and affects the inflammatory response to the oral pathogen *P. gingivalis* and subsequently caused infection (Muluke et al. 2016).

Gut microbiota, in conjunction with HFD, can influence the environmental factors affecting the progression of colorectal tumor development, such as the complex interrelationships with gut microbiota, inflammation, host genetics and can influence the development of CRC. There are many ways that microorganisms can induce chronic inflammation, including adherence to the epithelium, causing activation of an immune response through binding to Toll-like receptors and/or activating regulatory T (*Treg*) cells; synthesis and secretion of cytotoxic biomolecules or metabolites; or by translocation into the body (Greer and O'Keefe 2011; Vipperla and O'Keefe 2016).

The prior the diagnosis, the much easier is to control it. So at this very step machine learning can help to make a preliminary judgment about T2D on the basis of daily physical examination, and can serve as a reference for doctors (Alghamdi et al. 2017; Kavakiotis et al. 2017; Lee and Kim 2016). In machine learning (ML) methods, to choose the valid features and the correct classifier are the most important problems. Although recently numerous algorithms have been developed to predict diabetes. This includes the traditional ML methods (Kavakiotis et al. 2017), like support vector machine (SVM), decision tree (DT), logistic regression and many others in order to deal with large datasets (Razavian et al. 2015). ML methods are widely used in predicting diabetes and they get preferable results. Decision tree is one of favoured ML methods in medical sciences, which has great classification power. Random forest (RF) generates many decision trees. Therefore, in this study, we used decision tree and RF for prediction. The idea behind using ML methods for prediction studies is to make the call before it develops the disease.

In our previous study, it was shown that males and females of the different Collaborative Cross (CC) mouse lines varied significantly in T2D development and progression in response to diet challenges of CHD and HFD (Atamni et al. 2016a). Hence, we confirmed that sex and diet effects were prominent among the different CC lines and HFD was significantly correlated with inducing obesity and T2D as compared to CHD (Atamni et al. 2016a). However, it was found that different CC lines showed different susceptibility (Atamni et al. 2017; Atamni et al. 2016b). These studies have confirmed that sex and diet effects are among the complex traits controlled by multiple genetic factors of the CC population (Iraqi et al. 2014; Karkar et al. 2020).

Methods

Ethical statement

All the experiments and mouse usage described in this study were compatible with the standards for care and use of laboratory animals and approved by the Institutional Animal Care and Use Committee (IACUC) of Tel Aviv University (TAU), Israel (IACUC no. 01-19-013).

Study cohort

307 mice were used in this study from eight different CC lines, which have different genetic backgrounds; both sexes were represented in each study group. High molecular weight genomic DNA for the CC lines was initially genotyped using the mouse diversity array (MDA), which consists of 620 000 SNPs, and re-genotyped by mouse universal genotype array (MUGA-7500 markers) and eventually with MegaMuga (77 800 markers) SNP arrays to confirm their genotype status and variations in their genetic structure (Ogurtsova et al. 2017). These mice were maintained at our animal facility at TAU under suitable and agreed ethical conditions of temperature (21-23° C), humidity and daily supervision. Mice were weaned at the age of 3 weeks old, and then maintained separately by line and sex, with a maximum of five mice in an open-top cage and free access to water and rodent chow diet. Summary of the used mice in this study and their assignments in the different experimental groups are presented in (Table 1).

Study design

The experiments started when mice were 8 weeks old and lasted for a period of 12 weeks, with the mice on either a high-fat Dietary (HFD) challenge, with or without bacterial oral infection, or a standard chow diet (CHD), with or without bacterial oral infection. Body weight (g) was determined bi-weekly, and glucose tolerance ability was determined at the end time point of the experiment. Thus, the experiment included four experimental groups as detailed and named below. *CHD/no-infection*: This group was maintained on CHD for 12 weeks as a control group for the HFD group and treated with a placebo infection challenge (no bacterial infection). *HFD/no-infection*: This group was maintained on HFD for 12 weeks and treated with a placebo infection challenge as a control group for the bacterial infection challenge. *CHD/with-infection*: This group was maintained on CHD for 12 weeks and treated with a bacterial infection challenge as an experimental group for infection effect. *HFD/with-infection*: This group was maintained on HFD for 12 weeks and treated with bacterial infection challenge as an experimental group for the combined challenges of infection and HFD.

IP glucose tolerance test

To detect disturbances in glucose metabolism that can be linked to diabetes or pre-diabetic conditions. An intraperitoneally (IP) administered glucose load from the bloodstream, during 180 min following a glucose load. The glucose tolerance test measures the clearance ability. Following 6 hours (06:00-12:00 am), fasting with free access to water, fasting blood glucose levels were measured (time zero) and consequently a solution of glucose (2.5 mg glucose per gram body mass) was, intraperitoneally administered. Thereafter, blood glucose levels were measured at different time points (15, 30, 60, 120, and 180 minutes after glucose injection), using the Accu-Check Performa glucometer (AC PERFORMA KIT 53597 by Roche Ltd.) and glucose strips (AC Performa 50 F2 24049 by Roche Ltd.).

Dietary challenge

Mice were weaned at the age of 3 weeks and maintained until 8 weeks of age on a standard rodent chow diet (CHD; Altromin 1324 IRR, Altromin Spezialfutter GmbH & Co., Germany), which provides 11% Kcal from fat, 24% from protein, and 65% from carbohydrates. The dietary challenge started when the mice were 8 weeks old, when they were maintained on either CHD or HFD for the 12-week period of the experiment. The high-fat diet (HFD; TD.88137) was considered equivalent to a western diet, and was supplied by Teklad Global (Harlan Inc, Madison, WI, USA). The HFD provided 42.0% Kcal from fat, 15.3% from protein, and 42.7% from carbohydrates.

Bacterial culture

Porphyromonas gingivalis (Pg) strain ATCC 33277 and *Fusobacterium nucleatum* (Fn) s train P K 1 594 w ere g rown in p eptone yeast e xtract containing hemin and vitamin K (Wilkins Chalgren Broth, Oxoid Ltd, Basingstoke, UK), in an anaerobic chamber with 85% N2, 5% H2, and 10% CO2, followed by 3 washes of two minutes, each, in phosphate-buffered saline (PBS). The bacterial concentration was measured spectrophotometrically and standardized to OD650nm = 0.1 for Pg, corresponding to 1010 bacteria/ml; and OD660nm = 0.26 for Fn, corresponding to 109 bacteria/ml. Before the infection challenge, the two strains of bacteria were mixed together in a 1:1 (Pg:Fn) ratio.

Data Analysis

The IBM SPSS (statistical package for the social sciences) software platform Version 24 was used for data analysis. The variation between the CC lines and the significance ($P < .05$) was assessed One-way analysis of variance (ANOVA).

Area under the curve calculation

To assess the glucose tolerance ability status, area under the curve (AUC) of the IPGTT results were calculated according to the trapezoid role between time 0 and 180 minutes. This quantitative measurement of glucose clearance activity is made by the formula below:

AUC at time a−time b = (b min − a min)×(glucose levels at time a+b)/2 (1)

While total AUC is defined as AUC0−180, which
is =AUC0−15 +AUC15−30 +AUC30−60 +AUC60−120 +AUC120−180

Heritability and genetic coefficient of variation analysis in the CC lines

The heritability estimates were obtained from unpublished data of a wide variety of traits presently being studied at TAU on the same lines as of Iraqi et al. (2014). For this data, broad-sense heritability estimates (H^2) including epistatic but not dominance effects were obtained from a one-way ANOVA with CC lines as the main effect as follows:

$H2 = V_g/(V_g + V_e)$ (2)

where Ve is the environmental component of variance within lines = MSwithin V_g is the genetic component of variance among CC lines = $(MS_{between} - V_e)/n$,n is the average number of mice per line

Intestinal preparation for polyp counts

The intestines were collected and soaked in a small plate filled with phosphate buffered saline (PBS) until washed (510 minutes) to facilitate tissue use. Next, the small and large intestines were washed and washed with PBS at least twice. The small intestine was divided into three equal segments. Proximal, intermediate and distal. Each segment, including the colon, was cut horizontally with a razor blade, spread on 154 cm^2 Whatman paper, and fixed overnight with 10% neutral buffered formalin (NBF). Each length and width was measured with a ruler. The internal organs were washed with 70% ethanol, stained with 0.02% methylene blue (1.5 minutes on each paper) and held overnight in PBS on a shaker. The number of polyps in each intestinal segment, the length in cm, and the size in cm^2 (width x length) were recorded.

Genetic coefficient of variation

The absolute value of the genetic variation is obtained as the genetic standard deviation ($Vg^{0.5}$). The coefficient of variation (ratio of standard deviation to the mean) was used to measure the ratio of the genetic standard deviation to the mean, (also termed the genetic coefficient of variation, CV_G), as the comparable measure for unit-free evaluation of genetic dispersion (Garcia-Gonzalez et al. 2012; Houle 1992).

For TAU data, CVG was estimated as SDG =Mean (3)

where SD_G = the broad-sense genetic standard deviation among CC lines = $V_G^{0.5}$, Mean = mean trait value across all CC lines

Computational methods:

Classification models

In this study, we used decision tree (DT), naïve bayes (NaBa), k-nearest neighbors (KNN) and Random Forest (RF) as the classifiers. All classifiers were implemented using the Scikit-Learn Python package.

Decision tree

The decision tree model having a tree structure can describe the process of classification instances based on features (Salzberg 1994). In this study, we used Scikit-Learn's default implementation, with maximal tree depth of 10.

Naïve Bayes

The model of Naïve Bayes classification is based on a Bayesian network (BN) structure. A BN refers to graphical model for probability associations between a set of variables. In this study, we used Scikit-Learn's default gaussian Naïve Bayes.

K-Nearest Neighbors

In K-nearest neighbor (KNN) technique, the nearest neighbor is measured with respect to value of k that defines how many nearest neighbors need to be examined to describe the class of a sample point. In this study, we used scikit-learn's default implementation for K-Neighbors classifier (5 neighbors).

Random Forest

RF is a classification model using many decision trees. This algorithm was proposed by Breiman (Breiman 2001; Lin et al. 2014) as a multifunctional ML method performing the tasks of prediction and regression. In addition, RF is based on Bagging and it plays an important role in ensemble ML (Lin et al. 2014; Svetnik et al. 2003). It has been employed in several biomedical researches (Liao et al. 2016; Zhao et al. 2014). In this study, we used Scikit-Learn's default implementation, with 100 trees in a forest.

Regression models

We have used 2 regression models: linear regression and K-nearest neighbours regression, as they have produced some meaningful results for our data.

Model Validation

To evaluate the capability of the model, usually two validation methods; namely hold-out method and k-fold cross validation method are used (Kohavi 1995; Refaeilzadeh et al. 2016; Su et al. 2018). As per the goal of each problem and the size of data, we can choose a method of choice to solve the problem (Kim 2009). In this study, we used K(4)-fold cross-validation.

Results

The study presents an analysis of 307 mice generated from eight genetically different CC lines, with both sexes (156 male, 151 female). These mice were exposed to a dietary challenge (HFD vs. CHD) with or without oral co-infection challenge during a 12-week experimental period. The variations in the number of polyps within each sex after 12 weeks on either HFD or CHD under infection or non-infection conditions were determined, as well as the length and size of intestine.

Polyp number variation in different CC lines in response to dietary and infection challenge

Our observations show that there is a variation in polyp number between the different CC lines on HFD and CHD. To evaluate the effect of the HFD on the number of polyps developed, we compared the control group (CHD/no-infection) with the non-infected group on HFD (HFD/no-infection). Females showed

higher polyp number on CHD in infected condition. The effect of HFD was recognized in IL711 and IL3912 by comparing infected mice on HFD versus non- infected mice on HFD. Surprisingly, these two lines responded very differently to diet. In IL711, infected mice maintained on standard diet developed significantly ($P<0.05$) more polyps compared to infected mice at HFD with values of 9.80±1.2 and 6.80±0.86, respectively. In contrast, infected females of IL3912 maintained on standard diet developed significantly ($P<0.01$) less polyps (5.38±0.48) compared to infected mice on HFD (10.80±2.85) (Figure 1A).

Males of IL6018 were resistant to diet and infection challenges, while females of IL6018 on HFD without infection presented a significant variation ($P<0.01$) in polyp development compared to the other groups. In male mouse population, overall, we found a significant ($P<0.01$) effect of infection on males on CHD, where infected males on CHD developed less polyps (6.45±0.70) when compared to non-infected mice on CHD (9.18±1.13).

Furthermore, we found that there is a significant opposite effect ($P<0.05$) of HFD without infection, males maintained on HFD without infection developed less polyps (7.29±1.22) compared to mice on standard diet without infection (9.18±1.13). Contrarily, infected males maintained on HFD developed more polyps (8.34±1.17) when compared to infected males maintained on CHD (6.45±0.70) (Figure 1B).

Males of IL72, IL711 and IL3912 also showed a significant variation with $P<0.05$, $P<0.05$, $P<0.01$, respectively, between infected males on HFD and infected males maintained on CHD (Figure 1B). In both small and large intestines, separately, female mice on HFD developed more polyps compared to control mice (Figure 2A). In overall male mouse population, there were a significant variation ($P<0.01$) between non-infected males on CHD and infected males on CHD, where infected males developed less polyps in small intestines, 6.37±0.98 and 4.53±0.49, respectively. In addition, non-infected males on CHD developed significantly ($P<0.01$) more polyps compared to non-infected males on HFD, 6.37±0.98 and 4.64±0.95, respectively. (Fig. 2B). IL3912 and IL4141 males maintained on CHD without infection were, significantly different ($P<0.01$) from all the other 3 experimental groups as shown in whole intestines, as well.

Infected males of IL3348 maintained on HFD developed, significantly ($P<0.05$) more polyps (8.33±2.40) compared to infected males on CHD (5±0.45) (Figure 2B). The effect of the infection on polyp development, along the whole intestines was highly significant in the female population (Figure 3A). The effect of infection was observed in males of IL3912 and IL4141, since infected males on HFD developed significantly ($P<0.01$) less polyps in colon compared to non-infected males at CHD (Figure 3B). On the other hand, the combination of HFD and infection affected males of IL72 and IL711 and, significantly development of more polyps ($P<0.05$) in colon compared to infected mice maintained on (Figure 3B).

The combination of HFD and infection significantly affected ($P < .05$) the whole intestines and the small intestine in the female population compared to other groups. Overall, infected females maintained on HFD showed, significantly the shortest intestines with $P<0.01$ when compared to female mice on CHD and female mice at HFD without infection ($P<0.05$.) (Figure 4A), while in the overall male mouse

population, infected mice maintained on CHD showed, significantly longer intestines compared to the other three groups when compared to control male mice maintained on CHD ($P<0.05$) and when compared to male mouse maintained on HFD ($P<0.01$) (Figure 4B).

Non-infected females of IL6018 showed the longest small intestines (40.37±0.37 cm) significantly ($P<0.01$), compared to other three experimental groups of challenges (Figure 5A), while infected males of IL72 on CHD was significantly ($P<0.01$) shorter small intestines (25.5±0.47) compared to infected males maintained on HFD (30.78±0.39 cm) (Figure 5B). Overall infected male and female mice maintained on HFD presented shorter colons, 6.76± 0.49 cm and 6.53± 0.34 cm compared to other three groups of challenges (Figure 6A and 6B).

Combinatorial effect of the diet and infection on length and size of the intestine

Five tested different lines from the eight, showed significant variation between infected females maintained on HFD and infected females on CHD. Infected females maintained on HFD in all these lines showed shorter colon when compared to infected females on CHD.

The combinatorial effect of HFD and infection observed varied in most of the male lines. Similar to female response, infected males of IL72, IL557, IL3912, IL4141 and IL6018 maintained on HFD showed, significant short colon compared to infected males maintained on CHD.

In females' population (Figure7A): Overall infected females on CHD had, significant ($P<0.01$) the biggest size of intestines (36.15±1.35 cm^2) compared to non-infected females maintained on CHD (31.54±1.88 cm^2), Infected females at HFD (28.25±2.17 cm^2) and non-infected females at HFD (31.75±2.53 cm^2). Infected females of IL1912 and IL711 maintained on CHD also showed the biggest size of intestines compared to the other three experimental groups.

Results have showed that Infected females of IL557, IL711 and IL1912 on CHD developed significantly ($P<0.01$) bigger size of intestines, 39.21±0.1.08 cm^2, 40.06±0.3.5 cm^2, and 38.11±0.2.23 cm^2, respectively, when compared to infected females on HFD as showed records of 26.33±2.20 cm^2, 26.47±0.1.75 cm^2, and 26.36±4.59 cm^2, respectively.

The effect of oral infection on males was observed when mice maintained on CHD. Results have shown that infected males of IL557, IL1912 and IL3912 on CHD developed bigger size of intestines, with records of 43.36±3.26 cm^2, 37.45±1.49 cm^2 and 40.31±1.24 cm^2, respectively, when compared to non-infected males on CHD, which showed lower records, and were 32.37±0.1.50 cm^2, 26.95±1.58 cm^2 and 35.19±0.2.63 cm^2, respectively. However, infected males of IL72 on CHD responded, differently to infection and developed, significantly ($P<0.05$) smaller size of intestines (26.07±1.69 cm^2) compared to non-infected males of IL72 on CHD (31.07±0.63 cm^2) (Figure 7B)

Overall, infected females maintained on CHD significantly showed the biggest size of small intestines (29.67±1.24 cm^2) when compared to males maintained on CHD without infection, 25.11±1.68 cm^2 with

P<0.01, infected males on HFD, 23.81±1.94 cm² with P< 0.05 and non-infected males on HFD, 27.19±2.26 cm² with P<0.01.

Non-infected females of IL6018 and IL557 maintained on CHD, had a significant (P<0.05), bigger size of small intestines with records of 34.01±1.91 cm² and 33.14±1.29 cm², respectively, when compared to non-infected females maintained on HFD as showed records 25.74±1.56 cm² and 27.78±1.63 cm², respectively (Figure 8A)

Similarly, to overall females' mice response, infected males maintained on CHD showed significantly (P<0.01) the biggest size of small intestines (27.76±1.95 cm²) compared to non-infected males maintained on CHD as showed records 25.70±1.79 cm², to infected males maintained on HFD with records 24.22±1.51 cm², and to non-infected males when maintained on HFD with showed records 24.23±1.26 cm² (Figure 8B).

Overall males and females showed significant (P<0.01) bigger size of colon in CHD+No.inf and CHD+Inf compared to HFD+No.Inf and HFD+Inf experimental groups.

Non-infected females of IL557, IL3348, L4141 and IL6018 maintained on CHD showed significantly (P<0.01) bigger size of colon compared to non-infected females maintained on HFD (Figure 9A). Non-infected males of IL1912, IL3912 and IL6018 when maintained on CHD showed bigger size of colon with records of 6.90±0.31 cm², 8.18±1.15 cm² and 8.53±0.31 cm², respectively, compared to non-infected males maintained on HFD as showed records 2.58±0.54 cm², 5.05±0.37 cm² and 5.32±0.31 cm², respectively (Fig.9B). Another significant variation (P<0.01) was observed between non-infected males maintained on CHD and infected males on HFD in IL1912, IL3912 and IL6018 (Figure 9B).

Variation in the effect of HFD on glucose tolerance

The results presented in (Figure 10) revealed, significantly higher values of AUC compared to the control group (P < .01) in both male and female populations. The combination between HFD and infection induced high levels of AUC values, while male mice showed higher values of AUC (51805.95 ± 3483), when compared with males with males maintained on CHD with oral infection (37772.97± 2076.38) (Figure 10B).

The AUC profiles observed in females of IL72, IL4141 and IL3912, IL557 at HFD, which showed a significant difference comparing to females maintained on CHD of the same lines with levels of *P < 0.05 (Figure 10A). Interestingly, males of the same lines at HFD also show a significant variation in AUC (mg/dL*min) values compared to males at CHD (P <0.01).

Multimorbidity heatmaps of polyp counts, AUC, BW, length and size

A major aim of our proposed research was to study the effect of the host genetic background on diseases multimorbidity, to better understand the coexistence of multiple disease conditions in an individual.

Heatmaps were developed of these traits and searched for an association in development and severity between these traits. As presented in Figure 11, overall, the non-infected females and males when maintained on CHD (Figure 11A) showed negative correlation while infected mice maintained on CHD showed positive correlation between total AUC and length of intestines (Figure 11B), which indicates, to be more resistant to the effects of the dietary and infection challenges on the development of these traits.

This figure further analyzes the data showing which traits were found to be non-significant or significant at levels P < .05 and P < .01 in female and male mice of the studied lines in the different experimental groups. The results presented here, show that body weight gain was highly significantly (P < .01) increased in female IL557 mice compared with IL711 females in all experimental groups, while in male mice body weight gain was highly significant (P < .01) in all groups except the CHD + No-Inf. group. The same significance level (P < .01) was also observed in AUC values between females in the HFD + No-Inf. and CHD + No-Inf. group, and males in HFD + No-Inf. and in CHD + Inf groups. Intestine size was highly significantly different (P < .01) between IL557 and IL711 females in the CHD + No-Inf. group only, and between the male mice in the CHD + Inf group. In the case of the length of the intestines, a significant variation (P < .05) was observed between IL557 and IL711 females in the CHD + No-Inf. group only, and between male mice in two experimental groups, HFD + No-Inf. and CHD + Inf. Finally, the number of polyps was found to be a significant variant (P < .05) between the two lines only in HFD + No-Inf. Group, with no significant difference between sexes.

Heritability and genetic coefficient of variation

Table 2 presents heritability and genetic coefficient of variation values for a variety of traits studied. The heritability values of these traits are generally in the range of 0.33–0.92, while the CVG has been observed to be in the range 0.00–0.62, much higher than the benchmark of 0.071. Thus, the data shows that an absolute magnitude of genetic variation among the CC lines observed is higher than found within a typical outcrossing population.

Classification and Regression Results

The results presented in Table 3, the size or area of the intestine could be classified with high ROC AUC values for four lines. The model that performed best across the board was Random Forest with four values greater than 0.8 (lines 557, 711, 1912 and 3348) and a minimal value of 0.663 (line 6018). The single highest score was for Logistic Regression in line 557 – AUC = 0.936. In Supplementary Table 1, the classification of the mouse intestine length does not produce many high scores. The only line with multiple high values was 3912, for which the Logistic Regression performed best with a score of 0.859, while Naïve Bayes and Random Forest produced scores of 0.832 and 0.835 respectively. In lines 72 and 557 significantly lower scores – No model was able to predict with a score greater than 0.6.

The results presented in Supplementary Table 2 are the AUC classification scores also varying between different lines. While for most lines, we see no model with a ROC AUC greater than 0.8, in lines 72 and

1912 there are 2-3 models that produce such values. The best model for these lines was Logistic Regression with scores of 0.861 and 0.867. The observed results presented in Supplementary Table 3 are of the final body weight classified with exceptionally high AUC values for almost all lines. The model that performed best was Logistic Regression, with a maximum of 0.948 for line 4141 and a minimum of 0.682 for line 6018. The lines 72 and 6018 stand out as no model has produced a ROC AUC of over 0.8, while for most lines 3-5 models were able to produce such scores. These results are not surprising, as there is a strong positive correlation between initial body weight and final body weight in most lines. In Supplementary Table 4 a significant difference between two groups of lines: Lines IL72, IL711, IL1912 and IL3348 all had at least two models with high AUC values (0.823 and above), whereas lines IL557, IL3912, IL4141 and IL6018 had no model with an AUC of 0.75. The best model for this prediction was Random Forest with a peak of 0.985 for line 3348. In Supplementary Table 5 as we can see in the results, the number of polyps is very hard to classify by the chosen models. The results presented in Supplementary Table 6 indicate that diet can be classified with high AUC values for three lines using most models. For lines IL72 and IL4141 we were able to observe five models with AUC larger than 0.84. However, for some lines all models were unable to produce a high AUC value like the best result for line IL711 was 0.619.

The classification results are significantly different among different lines, models, and input-output combinations. While some features, such as polyp number were completely unpredictable, other features like final body weight are highly predictable with AUC values of ~0.9. Out of the 6 models we used, Logistic Regression, Naïve Bayes and Random Forest performed better than Decision Trees, Nearest Neighbors and SVC. Most lines obtained high AUC values for some combinations, yet line 6018 stands out as the significantly most unpredictable line, with only one AUC value larger than 0.8 – 0.801 for classifying the length using Random Forest.

Regression

In Supplementary Table 7, to predict the size of intestine, for most lines the behavior was completely unpredictable, with values of zero or close to zero. However, the linear models produced correlative predictions for line IL72 (median score of ~0.46) and for line IL1912 (median score of 0.229-0.299). The only line with correlative prediction was line IL6018. The Neighbors model performed best with an average score of 0.424 and a median score of 0.48. The linear models, Linear Regression and Lasso also produced correlative predictions with median scores of 0.299 and 0.336 respectively in predicting the length of intestine as presented in Supplementary Table 8.

The results presented in Supplementary Table 9 are the regression results for AUC varying drastically among the lines. For lines IL72, IL1912 and IL4141 both median and average results are very high, with a peak of 0.775 for Linear Regression in line IL72. On the contrary, the scores for the other 5 lines are virtually 0. In Supplementary Table 10, much like the high classification results for the final body weight, the numerical prediction for it is also very high. The peak was in Linear Regression for line IL4141 with a very high 0.877 median score. For 6 lines we were able to obtain correlative results in all models, and for

the remaining 2 lines 1-2 models produced weak correlative results – these results indicate some predictability. As the classification for the number of polyps produced the lowest predictability. The input combination of diet and infection status only produced higher results than any other subset of the original input. However, the results are still non-correlative as presented in Supplementary Table 11.

The Regression results vary drastically among models. Extra Trees Regression performed significantly worse than Nearest Neighbors and the 2 linear models – Linear Regression and Lasso. For most input-output combinations, we were able to observe the correlative results for at least 2 models indicating that these features are somewhat predictable. The number of polyps was completely unpredictable, with only one score larger than 0.2 in line IL1912. Most of lines obtained correlative results for some combinations, yet lines IL3348 and IL6018 stand out as the significantly most unpredictable lines, with only one predictable feature – final body weight and intestine length respectively.

Discussion

Diabetes mellitus is a disease, pre-ceded by prediabetic state, and the development of the disease is a continuous process. Although phenotyping is a strength of the study, as it allows us to observe the preliminary stages of the disease progression. How to exactly predict and diagnose this disease by using machine learning is worth studying being one of our main aims. We have taken some of the diabetes related traits like body weight, glucose tolerance ability at different time points, as our prediction tools, so it is quite worthy to use these methods as a strategy to predict and in turn prevent from T2D. It is a powerful tool to be used in our daily lives for not only prediction of diabetes but also its prevention by interfering within our lifestyle, diet and overall, our daily routine. The right diagnosis will lead us to introduce the prevention strategies through lifestyle, diet, exercise etc. to delay/avoid the development of T2D. It must be noted here that accurate prediction needs more indexes. In addition, by comparing the results of three classifications, we can find that there is little difference between RF and DT, but random forests are obviously better than other classifiers in some methods. The best result for the dataset is above 0.80 in most CC lines indicating ML can be used for prediction of diabetes, but finding suitable attributes, classifier and data mining methods are very important. According to our results, we usually chose the method using all features but still has a better result. LR and RF had the best result among the six classifiers as also observed by Zou et al. (2018). Therefore, our observations provide valuable insight into the potential application of the AUC as a predictive measure for T2D and highlight the need. These ML methods have also been recently applied by Ben-Assuli et al. (2022) for faster diagnosis and treatment of nonalcoholic fatty liver disease (NAFLD). In summary, our findings provide a clear indication that it can be used together with conventional risk factors, to predict multiple metabolic outcomes. Our results provide a significant resource for further studies to determine the causal relationship and the progression of T2D, therefore the prospect of using the personalized medicine is a promise. The results presented in this paper is a first step towards applying personalized/precision medicine approach based on early prediction and early prevention approaches.

It must also be noted that the successful classification methods rely on using environmentally affected traits, such as diet and body weight. However, the differences in the results between the lines indicate clearly that genetic background is key in attempts of such early diagnosis/prevention. There isn't a single model that predicts well for all genetic backgrounds, much like there isn't a single treatment that works for every genotype. The medicine should be personalized through the entire process – from diagnosis to treatment.

As one of the most prevalent diseases worldwide, T2D is a progressive disease, which makes up about 90% of cases of diabetes. In 2011, there were 280 million diabetic patients estimated to have diabetes, more than 500 million in to be diagnosed in 2030 (http://www.diabetesatlas.org/). It is a progressive condition in which the body's resistance to the normal effects of insulin increases greatly and/or eventually loses the capacity to produce enough insulin within the pancreas. Although T2D was shown to be a very complex disease with intertwined factors such as genetics, epigenetics and environmental, which are not yet fully understood in how they interact, other major environmental factors are well recognized and understood such as diet and activity level.

Insulin resistance compensated with hyperinsulinemia, typically characterizes the onset of T2D. A large spectrum of T2D susceptibility genes and many risk factors including obesity and a sedentary lifestyle were identified through extensive basic and clinical examinations, and these factors are shared with colon cancer development. One way to explain the diabetes-colon cancer factor relation is through the IGF-1 mechanism. The receptor ligand complex when activated by IGF-1, inhibits apoptosis, and allows progression through the cell cycle. Both colorectal epithelia and cancer cells express IGF-1 in vitro, hence showing that IGF-1 can influence both premalignant and cancerous stages in the cells. Similar to IGF-1, insulin stimulates growth of normal colonic and carcinoma cells in vitro (Björk et al. 1993) The mutagenic properties of insulin may be mediated through IGF-1 receptors, in contrast to its glycemic-control properties, despite the fact that colon cancer tissue has both insulin and IGF-1 receptors (Guo et al. 1992). Insulin increases bioactive IGF-1 through various mechanisms. The primary regulator for hepatic production of IGF-1 is the growth hormone, and insulin partly regulates the number of hepatic growth receptors (Giovannucci et al. 2001).

Accordingly, in our previous studies, we have shown the importance of the host genetic background. The immense role it played in determining the sensitivity of the mice to insulin resistance and body weight gain as well as the development of the polyps in intestine as well as other infectious diseases (Atamni et al. 2018; Shusterman et al. 2013a; Shusterman et al. 2013b; Lorè et al. 2015). Furthermore, a study from our lab reported the comorbidity between T2D and obesity (Iraqi et al. 2014).

In this project, we studied the effect of the host genetic background on the multimorbidity of T2D and intestinal cancer developments in response to diet and oral infection challenges. We have used the CC mice, which have been shown in previous studies to be a powerful platform for studying complex diseases.

Chronic inflammation was realized to be mainly caused by two important, yet most preventable factors, diet, and oral infection. These happen to be preventable causes of cancer and other chronic diseases including T2D. Oral microorganisms as proven in previous studies are associated with gastrointestinal cancers; and it is suggested evidently that specific bacterial infections promote the development of certain diseases (Zhang et al. 2019). In previous studies in different ethnic subgroups of human populations, a significant sex effect in T2D onset and progress, which differed between the males and females within and between different ethnic groups (Atamni et al. 2017; Atamni et al. 2016b; Gale and Gillespie 2001; Iraqi et al. 2014). Furthermore, our results in this study showed a significant sex effect on the number of polyps, length and size of intestines conducted in specific lines and under different experimental conditions of diet and infection. The glucose AUC which is an index of whole glucose excursion after glucose loading used for calculating the glycemic index. Our aim was to evaluate the possible usefulness of glucose AUC measurements in identifying cases of early stage of diabetes. Our data indicated that in overall population, males were more susceptible to glucose tolerance in comparison to females. This is explained by a longer time for clearing the blood glucose in IPGTT test subsequently showed higher values of AUC.

The males and females of IL6018 were resistant to the challenge and did not show any significant variation in their AUC values as a response to infection and diet. The HFD-fed females of IL6018 developed a significantly ($P<0.01$) higher number of polyps compared to other groups in the challenge. Previous studies have shown that the importance of host genetic background in the development of multiple polyps, while the larger number of polyps, the greater the chance that it possess cancerous cells (Grahn 2008). Focusing on the genetic factors, our ultimate aim is to be able to predict the intestinal cancer disease at early stages by identifying specific genes underlining the multimorbidity of intestinal cancer, obesity and T2D by using the CC model and its accompanied power of huge genetic diversity.

Multimorbidity, which is defined as the co-existence of two or more chronic conditions, is common in primary care patients, with at least fifty percent of patients over 50 years of age having two or more chronic conditions (Fortin et al. 2007). In order to investigate the link and the multimorbidity of intestinal cancer, obesity and T2D, our current study compromises a comprehensive picture of the susceptibility and the resistance. The diet and the infection separately, the heatmap analysis together with R measurements, and emphasize the strong correlation between these phenotypes. A significant correlation was observed between the total AUC Versus number of polyps, total AUC versus length of intestine and total AUC versus size of intestine in non-infected mice at CHD. The direction of each correlation varied between male and female population.

These findings highlight the impact of the added environmental factors to the previously present impact of the host's genetic background as the mice were largely exposed to the changes we manually increased in the environmental factors (infection, specifically). As we delved into deeper branches specifying the details within our results. The huge diversity between different strains under the same environmental conditions was demonstrated. Specifically in how multiple diseases developed simultaneously in each strain. This multimorbidity will help us to understand the mechanisms of each disease, and how it relates

to other diseases. Identifying one gene will help us to predict the development of many diseases and that gives us the opportunity to prevent it at early stages.

In conclusion the expected data promises to elucidate the nature of the genes involved in resistance and rate of development of intestinal cancer, T2D and obesity induced by HFD and oral infection to this multimorbidity. Once obtained, such data can be used to predict individual risk to develop these diseases and allow the development of a genetically based strategy for their prevention and treatment.

Declarations

ACKNOWLEDGMENTS

The authors declare no competing financial interests or other associations that may pose a conflict of interest (e.g., pharmaceutical stock ownership, consultancy). This report was supported by Binational Science Foundation (BSF) grant number 2015077, German Israeli Science Foundation (GIF) grant I-63-410.20-2017, Israeli Science Foundation (ISF) grant 1085/18 and core fund form Tel-Aviv University.

CONFLICT OF INTEREST

None.

References

1. Atamni HJ, Nashef A, Iraqi FA (2018) The Collaborative Cross mouse model for dissecting genetic susceptibility to infectious diseases. Mamm Genome 29:471–487.
2. Atamni HJ, Mahmoud E, Yaser S, Aysar N, Iraqi , FA (2016b). The Collaborative Cross mouse genetic reference population designed for dissecting complex traits. Chinese Journal of Comparative Medicine 26:1-19.
3. Atamni HJ, Ziner Y, Mott R et al. (2017) Glucose tolerance female-specific QTL mapped in collaborative cross mice. Mamm Genome 28:20–30.
4. Alghamdi M, Al-Mallah M, Keteyian S, Brawner C, Ehrman J, Sakr S (2017) Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. PLoS ONE 12(7): e0179805. https://doi.org/10.1371/journal.pone.0179805
5. American Cancer Society (2014) Explore Research, Cancer Facts and Statistics: Cancer Facts & Figures; 2014.
6. Atamni HJ, Mott R, Soller M, Iraqi, FA (2016a) High-fat-diet induced development of increased fasting glucose levels and impaired response to intraperitoneal glucose challenge in the collaborative cross mouse genetic reference population. BMC Genetics 17:10.
7. Atanasova KR, Yilmaz O (2014) Looking in the Porphyromonas gingivalis cabinet of curiosities: the microbium, the host and cancer association. Mol Oral Microbiol 29:55–66, 2014.

8. Ben-Assuli O, Jacobi A, Goldman O, Shenhar-Tsarfaty S, Rogowski O, Zeltser D, Shapira I, Berliner S, Zelber-Sagi S (2022) Stratifying Individuals into Non-Alcoholic Fatty Liver Disease Risk Levels using Time Series Machine Learning Models. Journal of Biomedical Informatics doi: https://doi.org/10.1016/j.jbi.2022.103986
9. Björk J, Nilsson J, Hultcrantz R, Johansson C (1993) Growth-regulatory effects of sensory neuropeptides, epidermal growth factor, insulin, and somatostatin on the non-transformed intestinal epithelial cell line IEC-6 and the colon cancer cell line HT 29. Scand J Gastroenterol 28(10):879-84.
10. Blasco-Baque V, Garidou L, Pomié C et al. (2017) Periodontitis induced by Porphyromonas gingivalis drives periodontal microbiota dysbiosis and insulin resistance via an impaired adaptive immune response. Gut 66(5):872-885.
11. Breiman L (2001) Random Forests. Machine Learning, 45:5-32. http://dx.doi.org/10.1023/A:1010933404324
12. Campbell, PT et al. (2010) Prospective study reveals associations between colorectal cancer and type 2 diabetes mellitus or insulin use in men. Gastroenterology 139:1138–1146.
13. Cavicchia PP et al. (2013) Racial disparities in colorectal cancer incidence by type 2 diabetes mellitus status. Cancer Causes Control 24:277–285.
14. Defronzo, RA (2009) From the triumvirate to the ominous octet: A new paradigm for the treatment of type 2 diabetes mellitus. in Diabetes 58(4): 773-795.
15. Fortin M, Soubhi H, Hudon C, Bayliss EA, van den Akker M (2007) Multimorbidity's many challenges. BMJ 334(7602):1016-1017.
16. Gale E, Gillespie K (2001) Diabetes and gender. Diabetologia 44:3–15.
17. Gallagher EJ, LeRoith D (2010) The proliferating role of insulin and insulin-like growth factors in cancer. Trends in endocrinology and metabolism: TEM 21(10) 610–618.
18. Garcia-Gonzalez F, Simmons LW, Tomkins JL, Kotiaho JS, Evans JP (2012) Comparing evolvabilities: common errors surrounding the calculation and use of coefficients of additive genetic variation. Evolution: International Journal of Organic Evolution 66(8):2341-9.
19. Giovannucci E (2001) Insulin, Insulin-Like Growth Factors and Colon Cancer: A Review of the Evidence, *The Journal of Nutrition*, Volume 131, Issue(11), Pages :3109S–3120S.
20. Giovannucci E (1995) Insulin and colon cancer. Cancer Causes Control 6:164–179.
21. Grahn, Sarah W, Madhulika GV(2008) “Factors that increase risk of colon polyps.” Clinics in colon and rectal surgery 21(4): 247-55.
22. Greer JB, O'Keefe SJ (2011) Microbial induction of immunity, inflammation, and cancer. Front Physiol 26; 1:168.
23. Guo, YS, Narayan, S, Yallampalli, C, Singh, P (1992) Characterization of insulin-like growth factor I receptors in human colon cancer. Gastroenterology 102:1101–1108.
24. Houle D (1992) Comparing evolvability and variability of quantitative traits. Genetics 130(1):195-204.

25. Iraqi AFA, Athamni HJ, Dorman A, Salymah Y, Tomlinson I, Shusterman A, Weiss E, Houri-Haddad Y, Mott R, et al. (2014) Heritability and coefficient of genetic variation analyses of phenotypic traits provided strong basis for high-resolution QTL mapping in the Collaborative Cross mouse reference population. Mamm Genome 25(3):109-119.
26. Iyengar NM, Hudis CA, Dannenberg AJ (2015) Obesity and cancer: local and systemic mechanisms. Annul Rev Med 66:297–309.
27. Jiang Y, Pan Y, Rhea PR, Tan L, Gagea M, Cohen L, Fischer SM, Yang P (2016) A sucrose-enriched diet promotes tumorigenesis in mammary gland in part through the 12-lipoxygenase pathway. Cancer Res 76:24–9.
28. Karkar L, Atamni, H, Milhem A, Houri-Haddad Y, Iraqi FA (2020) Assessing the host genetic background effects on type 2 diabetes and obesity development in response to mixed-oral bacteria and high-fat diet using the collaborative cross mouse model. Animal models and experimental medicine 3(2):152–159.
29. Kasuga M (2006) Insulin resistance and pancreatic beta cell failure. The Journal of clinical investigation, 116(7):1756–1760.
30. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I (2017) Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal 15:104-16. doi: 10.1016/j.csbj.2016.12.005
31. Kim JH (2009) Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Computational statistics & data analysis 53(11):3735-45. doi: 10.1016/j.csda.2009.04.009
32. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. InIjcai 14:1137-1145.
33. Kolb H, Martin S (2017) Environmental/lifestyle factors in the pathogenesis and prevention of type 2 diabetes. BMC medicine 15(1):131.
34. Lee BJ, Kim JY (2016) Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. IEEE journal of biomedical and health informatics 20(1):39-46. doi: 10.1109/JBHI.2015.2396520
35. Liao Z, Ju Y, Zou Q (2016) Prediction of G protein-coupled receptors with SVM-prot features and random forest. Scientifica 2016. doi: 10.1155/2016/8309253
36. Lin C, Chen W, Qiu C, Wu Y, Krishnan S, Zou Q (2014) LibD3C: ensemble classifiers with a clustering and dynamic selection strategy. Neurocomputing 123:424-35. doi: 10.1016/j.neucom.2013.08.004
37. Lone MI, Iraqi FA (2022) Genetics of murine type 2 diabetes and comorbidities. Mamm Genome. 3:1-6.
38. Lorè NI, Iraqi FA, Bragonzi A (2015) Host genetic diversity influences the severity of Pseudomonas aeruginosa pneumonia in the Collaborative Cross mice. BMC Genet 16:106.
39. Muluke M, Gold T, Kiefhaber K, Al-Sahli A, Celenti R, Jiang H, Cremers S, Van Dyke T, Schulze-Späte U (2016) Diet-Induced Obesity and Its Differential Impact on Periodontal Bone Loss. Journal of dental

research 95(2):223–229.

40. Ogurtsova K et al. (2017) IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040. Diabetes Res Clin. Pract 128:40-50.
41. Peeters PJ, Bazelier MT, Leufkens HG, de Vries F, De Bruin ML (2015) The risk of colorectal cancer in patients with type 2 diabetes: associations with treatment stage and obesity. Diabetes Care 38:495–502.
42. Rakoff-Nahoum S, Medzhitov R (2009). Toll-like receptors and cancer. Nat Rev Cancer 9:57–63.
43. Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D (2015) Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. Big Data 3(4):277-87. doi: 10.1089/big.2015.0020
44. Refaeilzadeh P, Tang L, Liu H (2016) "Cross-validation," in Encyclopedia of Database Systems, eds L. Liu and M. T. Özsu (New York, NY: Springer):532–538.
45. Renehan AG, Zwahlen M, Egger M. (2015) Adiposity and cancer risk: new mechanistic insights from epidemiology. Nat Rev Cancer 15:484–98.
46. Salzberg SL (1994) by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Machine Learning 1:6.
47. Shusterman A, Durrant C, Mott R, Schaefer A, Weiss EI, Iraqi FA, Houri-Haddad Y (2013b) Host susceptibility to periodontitis: Mapping murine genomic regions. Journal of Dental Research 92:438-443.
48. Shusterman A, Salyma Y, Nashef A, Soller M, Wilensky A, Mott R, Weiss EI, Houri-Haddad Y, Iraqi FA. (2013a) Genotype is an important determinant factor of host susceptibility to periodontitis in the Collaborative Cross and inbred mouse populations. BMC Genet 9:14:68.
49. Su ZD, Huang Y, Zhang ZY, Zhao YW, Wang D, Chen W, Chou KC, Lin H (2018) iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. Bioinformatics 34(24):4196-204. doi: 10.1093/bioinformatics/bty508.
50. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of chemical information and computer sciences 43(6):1947-58. doi: 10.1021/ci034160g
51. Szczepanski MJ, Czystowska M, Szajnik M, Harasymczuk M, Boyiadzis M, Kruk-Zagajewska A, Szyfter W, Zeromski J, Whiteside TL (2009) Triggering of Toll-like receptor 4 expressed on human head and neck squamous cell carcinoma promotes tumor development and protects the tumor from immune attack. Cancer Res 69(7):3105-13.
52. Tabak AG et al. (2009) Trajectories of glycaemia, insulin sensitivity, and insulin secretion before diagnosis of type 2 diabetes: an analysis from the Whitehall II study. Lancet 373:2215–2221.
53. Tan J, Chen YX (2016) Dietary and Lifestyle Factors Associated with Colorectal Cancer Risk and Interactions with Microbiota: Fiber, Red or Processed Meat and Alcoholic Drinks. Gastrointestinal tumors, 3(1):17–24.

54. Tang H, Zhao YW, Zou P, Zhang CM, Chen R, Huang P, Lin H (2018) HBPred: a tool to identify growth hormone-binding proteins. International Journal of Biological Sciences 14(8):957. doi: 10.7150/ijbs.24174
55. The American Cancer Society (2020) medical and editorial content team. Key statistics for colorectal cancer. American cancer society. Cancer.org 1.800.227.2345.
56. Vipperla K, O'Keefe SJ (2016) Diet, microbiota, and dysbiosis: a 'recipe' for colorectal cancer. Food Funct 7(4):1731-40.
57. Whitmore SE, Lamont RJ (2014) Oral Bacteria and Cancer. PLOS Pathogens 10(3): e1003933.
58. Zhang Y, Niu Q, Fan W, Huang F, He H (2019) Oral microbiota and gastrointestinal cancer. OncoTargets and therapy 12:4721Zhang, Yangyang et al. “Oral microbiota and gastrointestinal cancer.” OncoTargets and therapy 12:4721-4728..
59. Zhao X, Zou Q, Liu B, Liu X (2014) Exploratory predicting protein folding model with random forest and hybrid features. Current Proteomics 11(4):289-99. doi: 10.2174/15701646110415012111515 4
60. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H (2018) Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics 515. https://doi.org/10.3389/fgene.2018.00515

Tables

Table 1. Summary of the assessed experimental mice of the 8 different CC lines and their assignments in each experimental group of challenge. Overall, 307 CC mice from 8 different CC lines, maintained either on HFD (42% fat) or CHD (18% fat) for 12 weeks, while half of them in each diet were infected at week 5 of the experiment. Males and females were exposure to the same conditions of diet and infection (151 females and 156 males).

	HF diet				CH diet				
	(+) .Inf		(-) .Inf		(+) .Inf		(-) .Inf		
CC Line	♀	♂	♀	♂	♀	♂	♀	♂	Total
IL72	3	7	3	3	4	4	2	2	28
IL557	6	3	5	5	3	6	5	3	36
IL711	5	4	5	5	5	5	5	4	38
IL1912	5	5	3	4	6	6	10	10	49
IL3348	5	3	2	3	3	6	4	3	29
IL3912	5	4	3	5	8	7	4	5	41
IL4141	4	2	4	2	6	3	4	8	33
IL6018	5	8	4	8	11	8	4	5	53
									307

Table 2: Summary of heritability and genetic coefficient of variation in the used CC lines in our study. Each experimental group includes CHD-non-infection, CHD-infection, HFD-non-infection and HFD-infection.

Sex	Trait	H2				CVg			
		CHD		HFD		CHD		HFD	
Male		Non inf	Inf	Non inf	Inf	Non inf	Inf	Non inf	Inf
	Total AUC	0.806	0.793	0.926	0.297	0.330	0.302	0.503	0.176
	BW	0.697	0.753	0.733	0.688	0.142	0.169	0.176	0.144
	NOP	0.332	0.483	0.370	0.023	0.195	0.299	0.313	0.061
	SIZE	0.408	0.293	0.638	0.191	0.124	0.125	0.191	0.089
	Length	0.540	0.673	0.716	0.498	0.095	0.123	0.113	0.082
Female	**Total AUC**	0.777	0.693	0.718	0.645	0.0001	0.175	0.285	0.247
	BW	0.007	0.790	0.529	0.445	0.009	0.142	0.187	0.162
	NOP	0.487	0.007	0.281	0.218	0.626	0.028	0.213	0.171
	SIZE	0.754	0.707	0.231	0.176	0.212	0.240	0.092	0.084
	Length	0.728	0.462	0.606	0.524	0.168	0.091	0.128	0.119

Table 3: Classification results. The Input features: sex, diet, infection, initial body weight and number of polyps. The Output: classification of size/area of intestine – larger/smaller than the 80th percentile of the data. Using the sex, diet, infection, initial body weight and number of polyps to classify whether size/area of intestine would be in the top 20% of the data has produced high values for most lines.

	72	557	711	1912	3348	3912	4141	6018
N	22	30	36	40	29	40	28	49
DT	0.566	0.73	0.705	0.697	0.786	0.645	0.631	0.572
NaBa	0.657	**0.882**	0.584	0.774	**0.842**	0.73	0.784	0.782
KNN	0.557	0.751	0.744	0.745	0.659	0.634	0.7	0.61
RF	0.76	**0.873**	**0.822**	**0.806**	**0.809**	0.723	0.765	0.663
SVC	0.357	0.347	0.385	0.455	0.356	0.387	0.404	0.381
LR	0.705	**0.936**	0.643	0.69	**0.806**	0.672	0.677	0.699

Figures

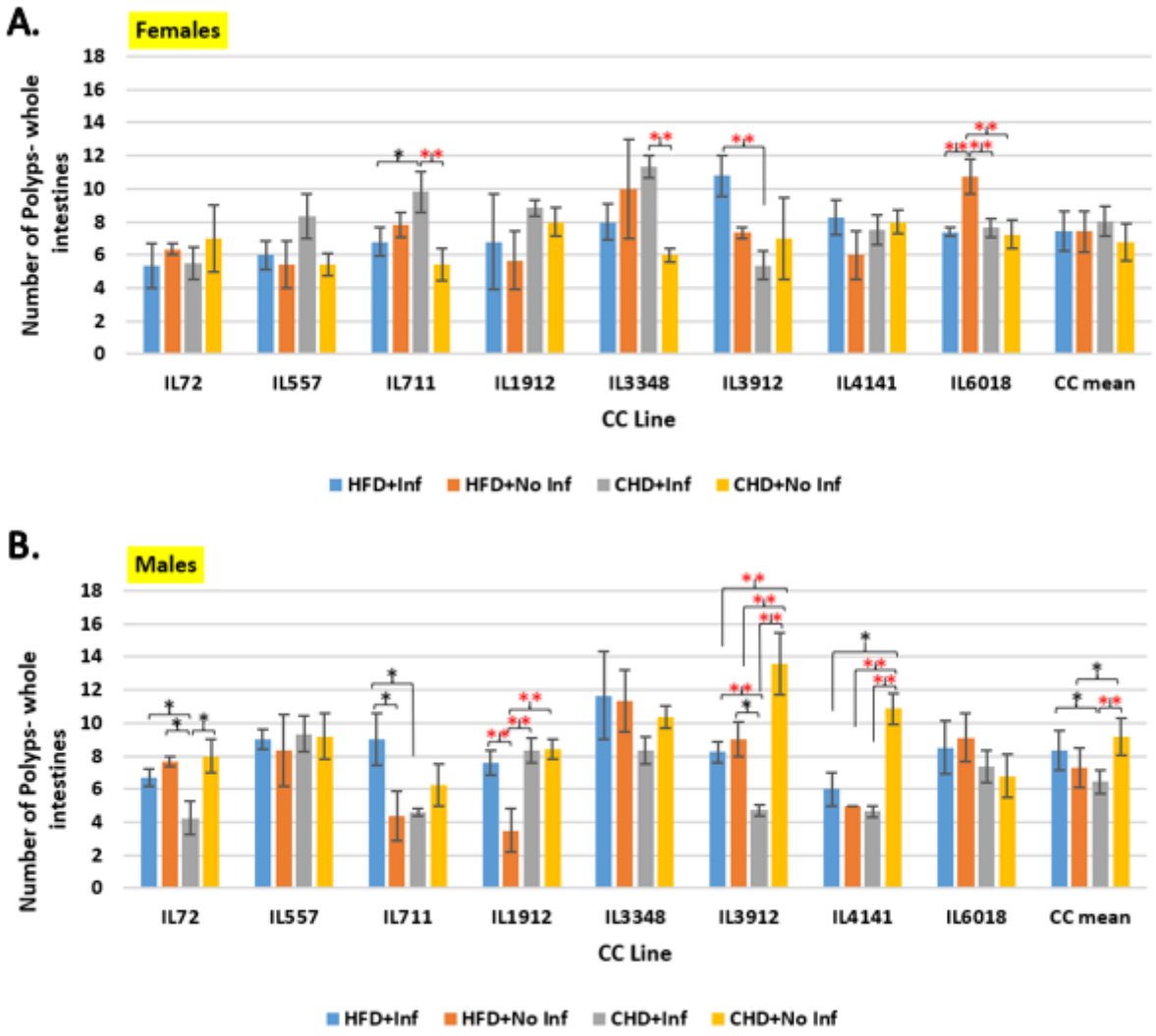


Figure 1

Number of polyps in the whole intestine sections (small and large) in male and female mice of eight different CC lines following 12 weeks of maintaining on either standard chow diet (CHD) or high-fat (42% Fat) diet (HFD) and with or oral infection challenges. X-axis presents CC lines and the mean of the entire population and Y-axis presents number of polyps in whole intestines in female [A] and male [B] mice.

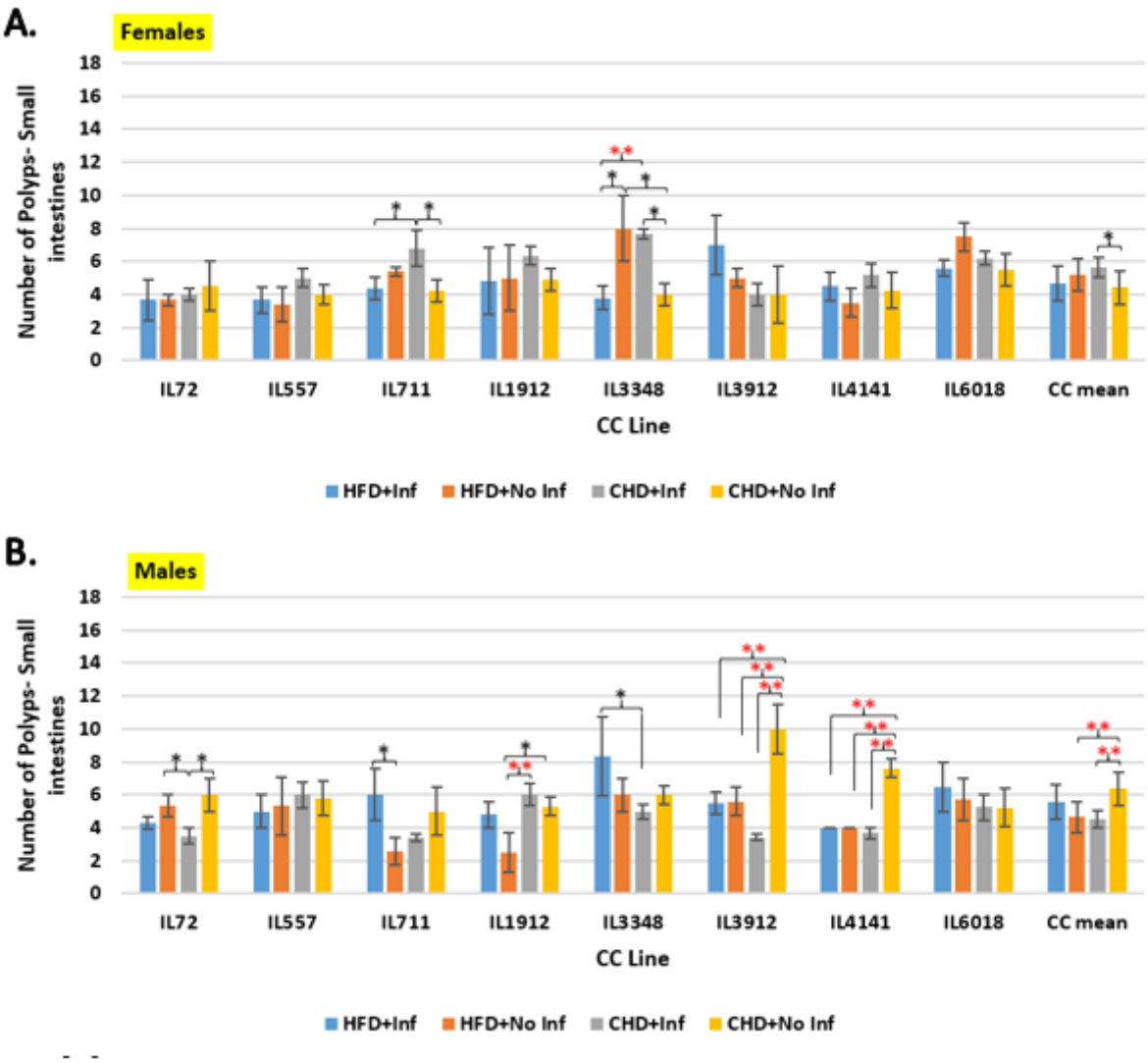


Figure 2

Number of polyps in small intestine of male and female mice of eight different CC lines and mean population after maintained for 12 weeks either on standard chow (18% Fat) diet (CHD) or high-fat (42% Fat) diet (HFD), and with or without oral infection challenges. X-axis presents CC lines and the mean of the entire population; Y-axis presents number of polyps in small intestine in female [A] and male [B] mice.

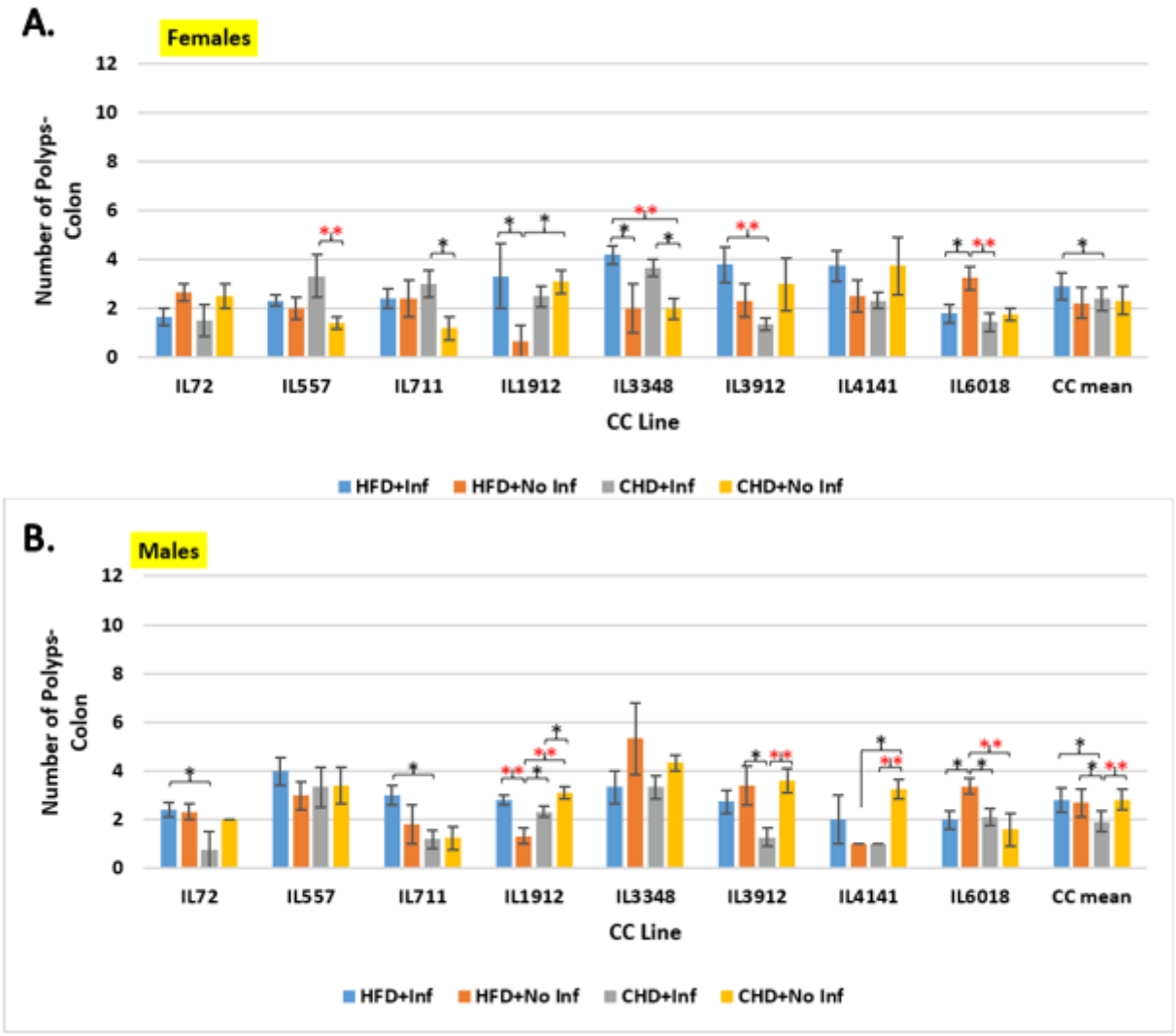


Figure 3

Number of polyps in colon in male and female mice of eight different CC lines and mean of the whole population after maintained for 12 weeks either on standard chow (18% Fat) diet (CHD) or high-fat (42% Fat) diet (HFD), and with or without oral infection challenges. X-axis presents CC lines and the mean of the entire population; Y-axis presents number of polyps in colon in female [A] and male [B] mice.

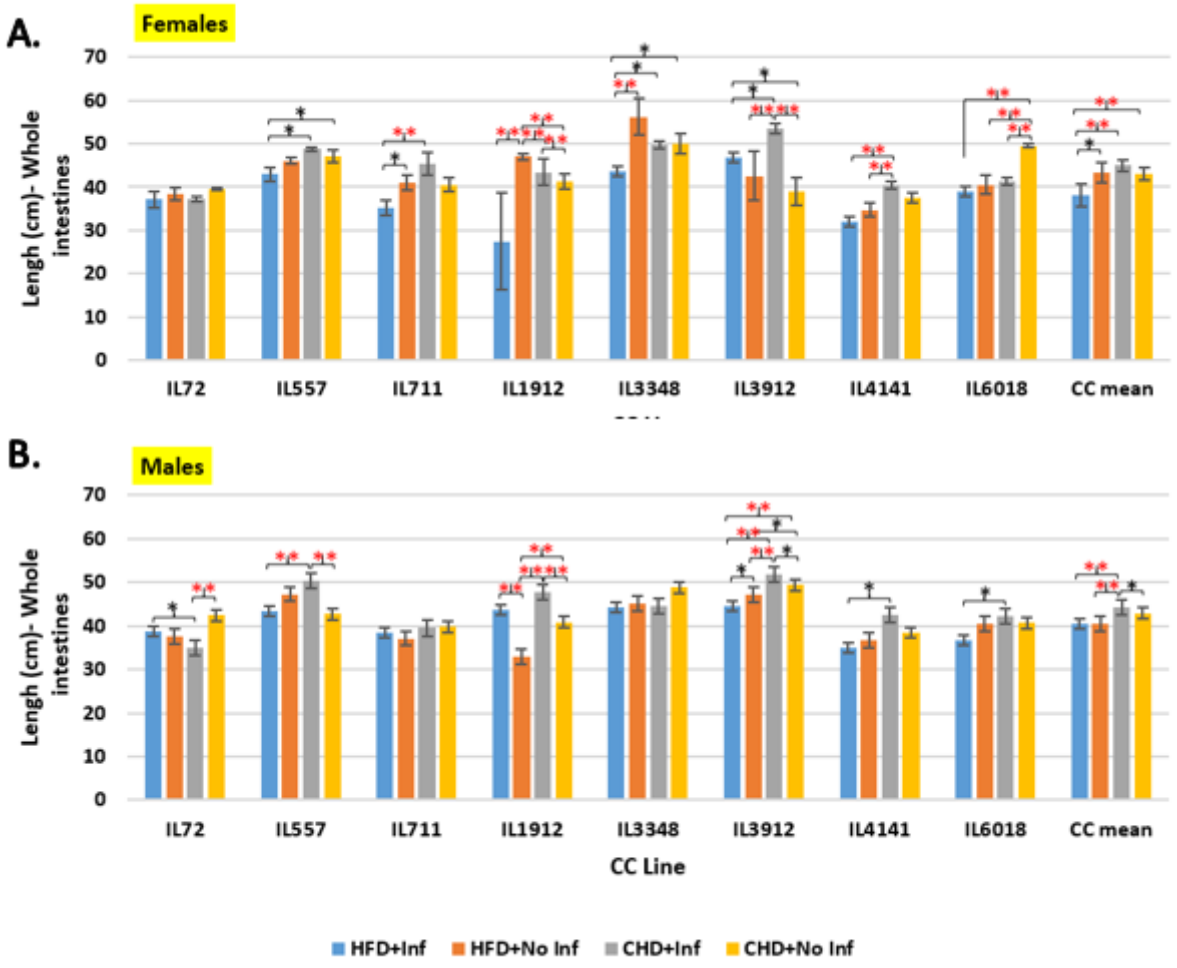


Figure 4

Length of whole intestines in cm of male and female mice of eight different CC lines and mean of the whole population after maintained for 12 weeks either on standard chow diet (CHD) or high-fat (42% Fat) diet (HFD) and with or without oral infection challenges. X-axis presents CC lines and mean of the whole population; Y-axis presents length of whole intestines (cm) in female [A] and male [B] mice.

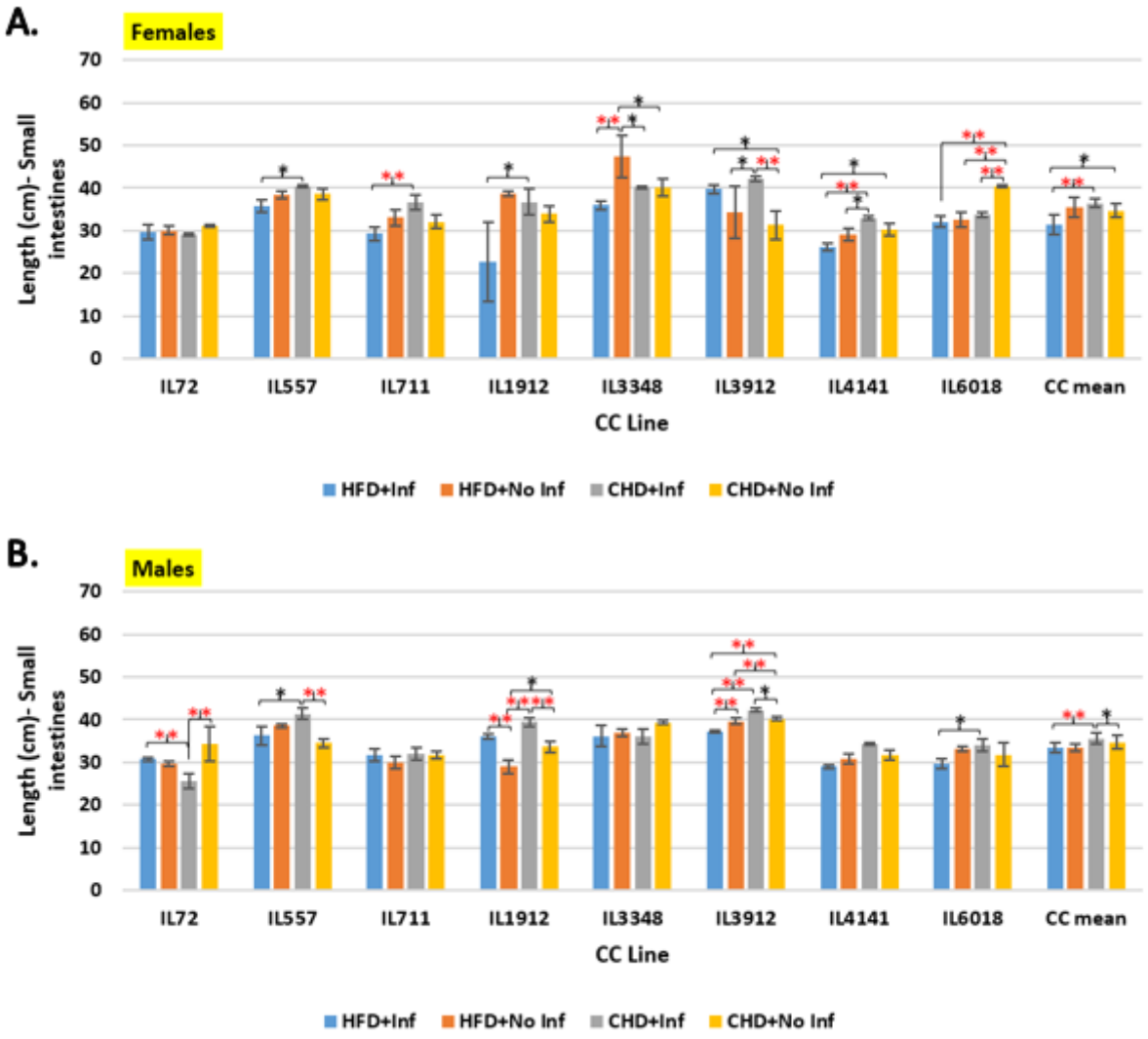


Figure 5

Length of small intestines in cm of male and female mice of eight different CC lines and their mean population after maintained for 12 weeks either on standard chow diet (CHD) or high-fat (42% Fat) diet (HFD) and either with or without oral infection challenges. X-axis presents CC lines and mean of the whole population; Y-axis presents length of small intestines (cm) in small intestines in female [A] and male [B] mice.

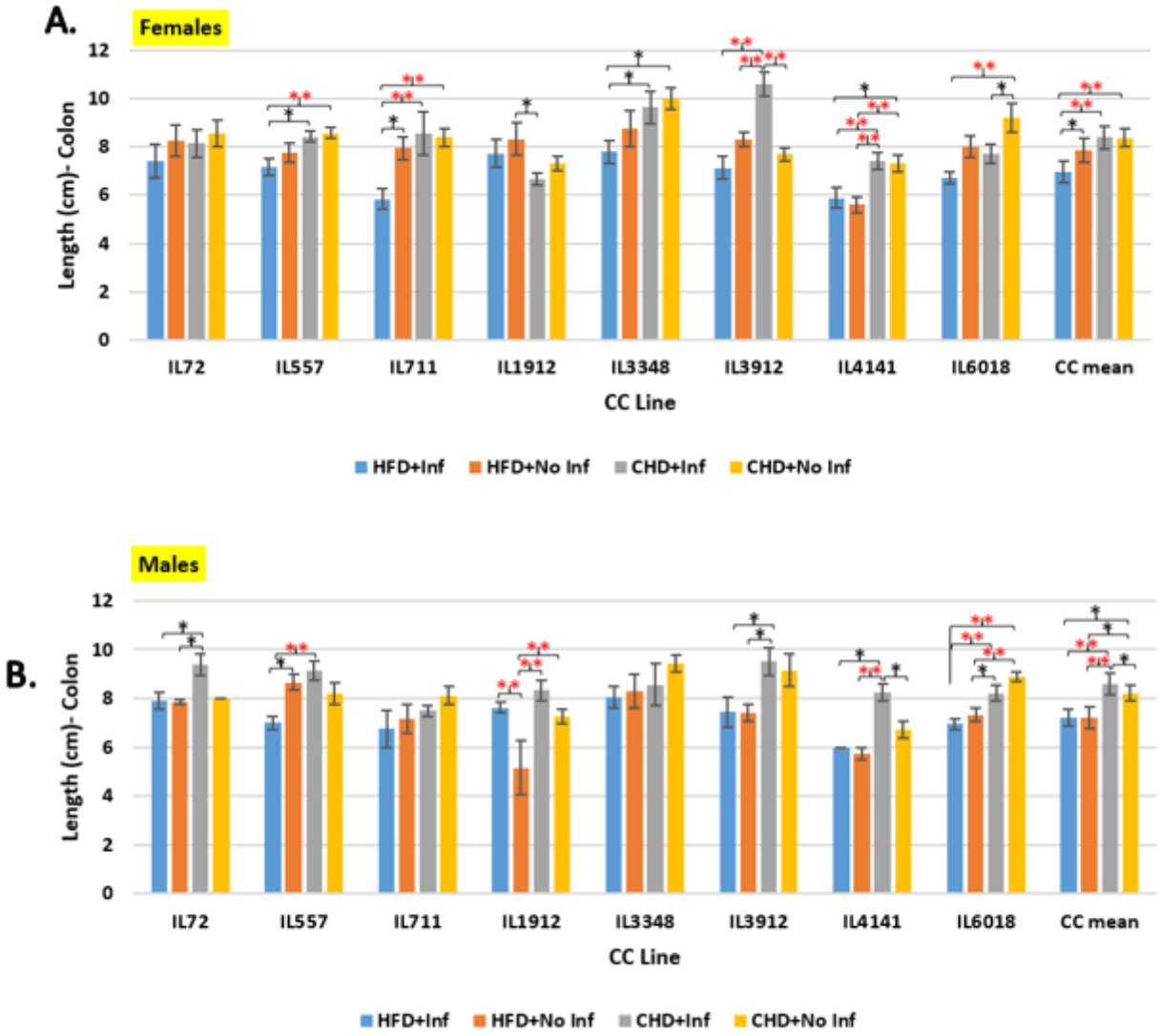


Figure 6

Length of colon in cm of male and female mice of eight different CC lines and their mean population after maintained for 12 weeks on either standard chow (18% Fat) diet (CHD) or high-fat (42% Fat) diet (HFD), and with or without oral infection challenges. X-axis presents CC lines and their mean of the entire population; Y-axis presents lengths of colon in cm in female [A] and male [B] mice.

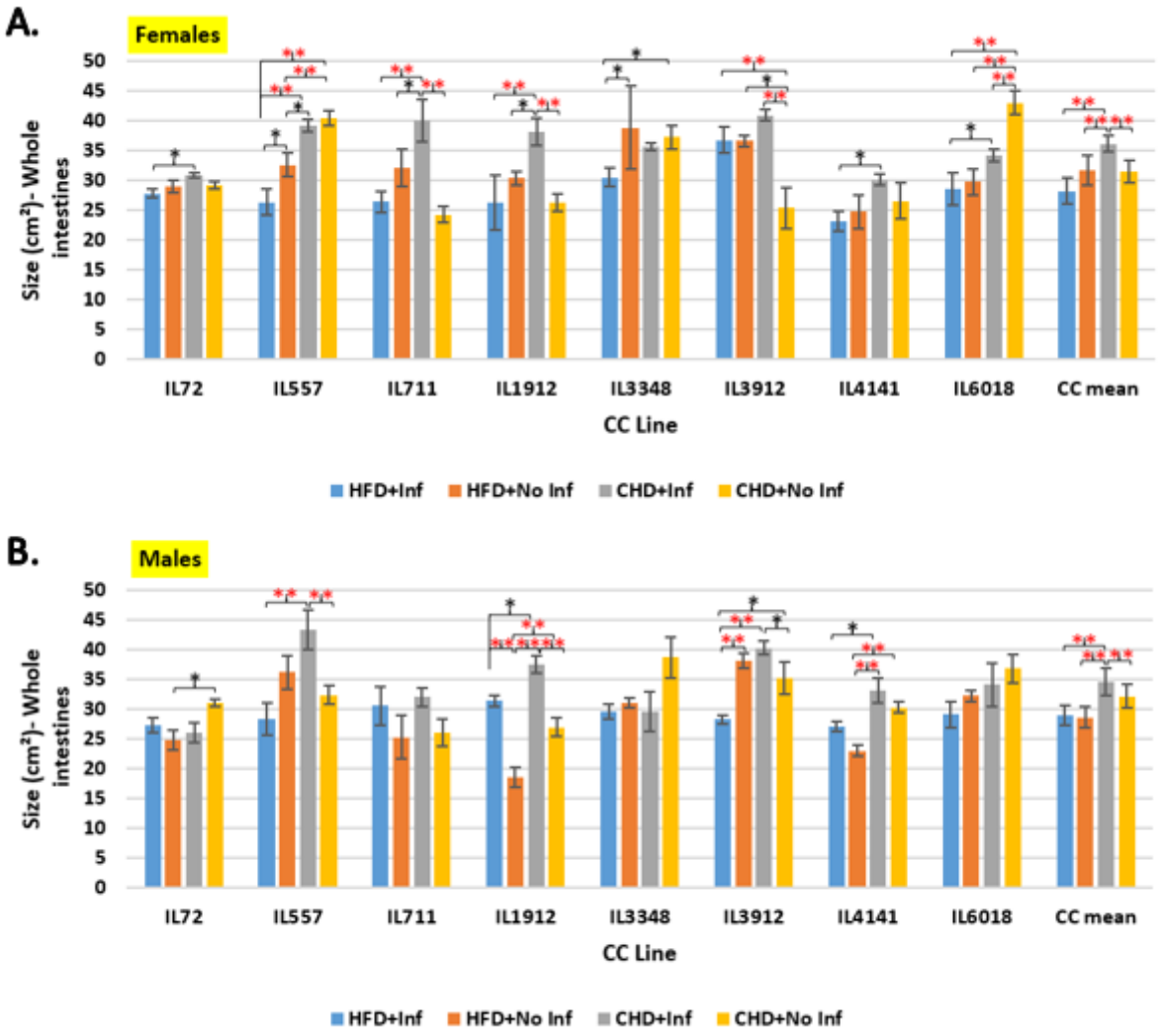


Figure 7

Size of whole intestines in cm^2 of male and female mice of eight different CC lines and mean population after maintained for 12 weeks on either standard chow diet (CHD) or high-fat (42% Fat) diet (HFD) and either with or without oral infection challenged. X-axis presents CC lines and the mean of the entire population; Y-axis presents size of the whole intestine in cm^2. Figure A shows data of female and B for male mice.

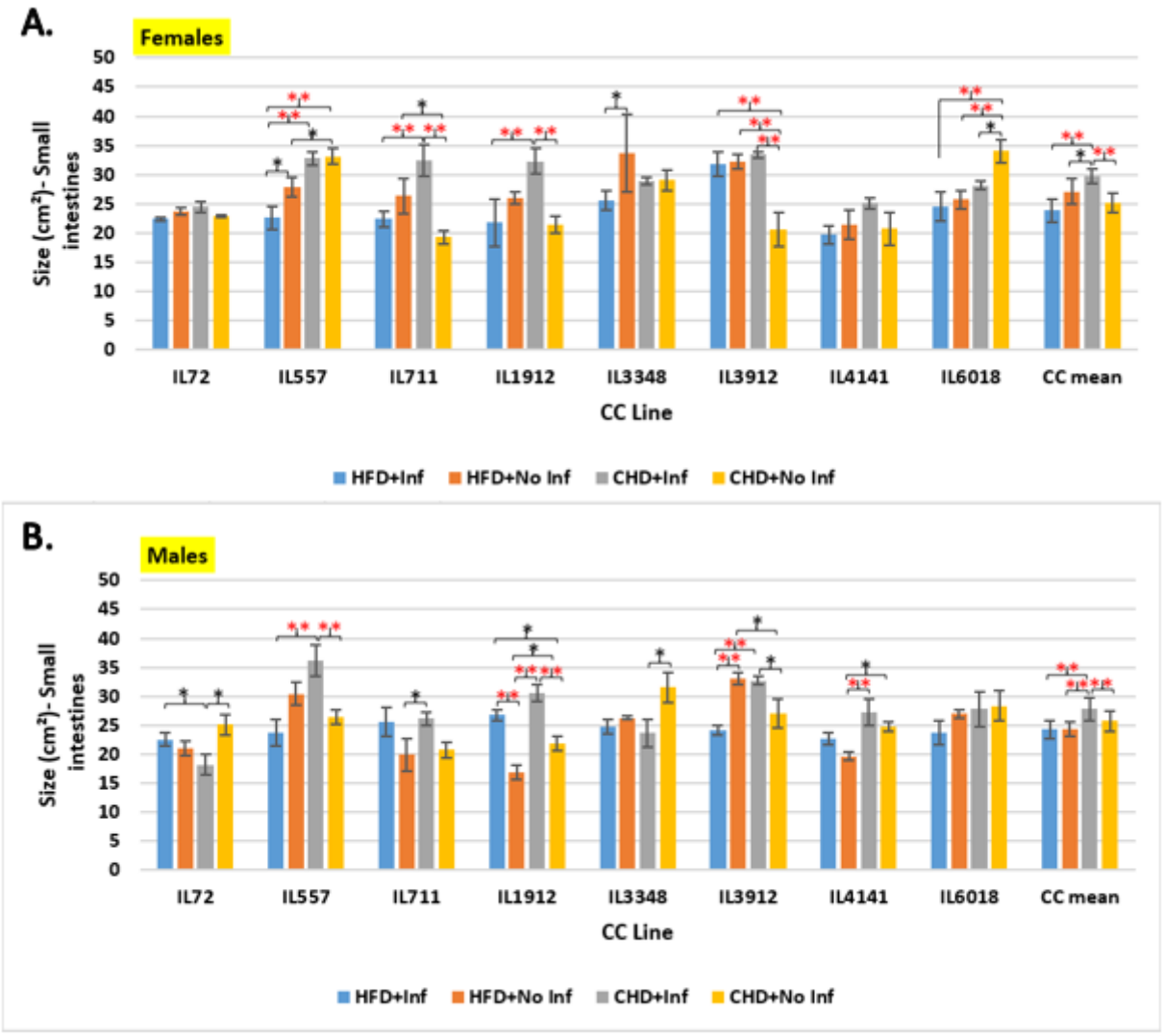


Figure 8

Size of small intestines in cm² of female and male mice of eight different CC lines and mean population after were maintained for 12 weeks either on chow (18% Fat) diet or high-fat (42% Fat) diet (HFD) and with or without oral infection. X-axis presents CC lines and mean of the entire population; Y-axis presents values of size of the small intestines in female [A] and male [B] mice.

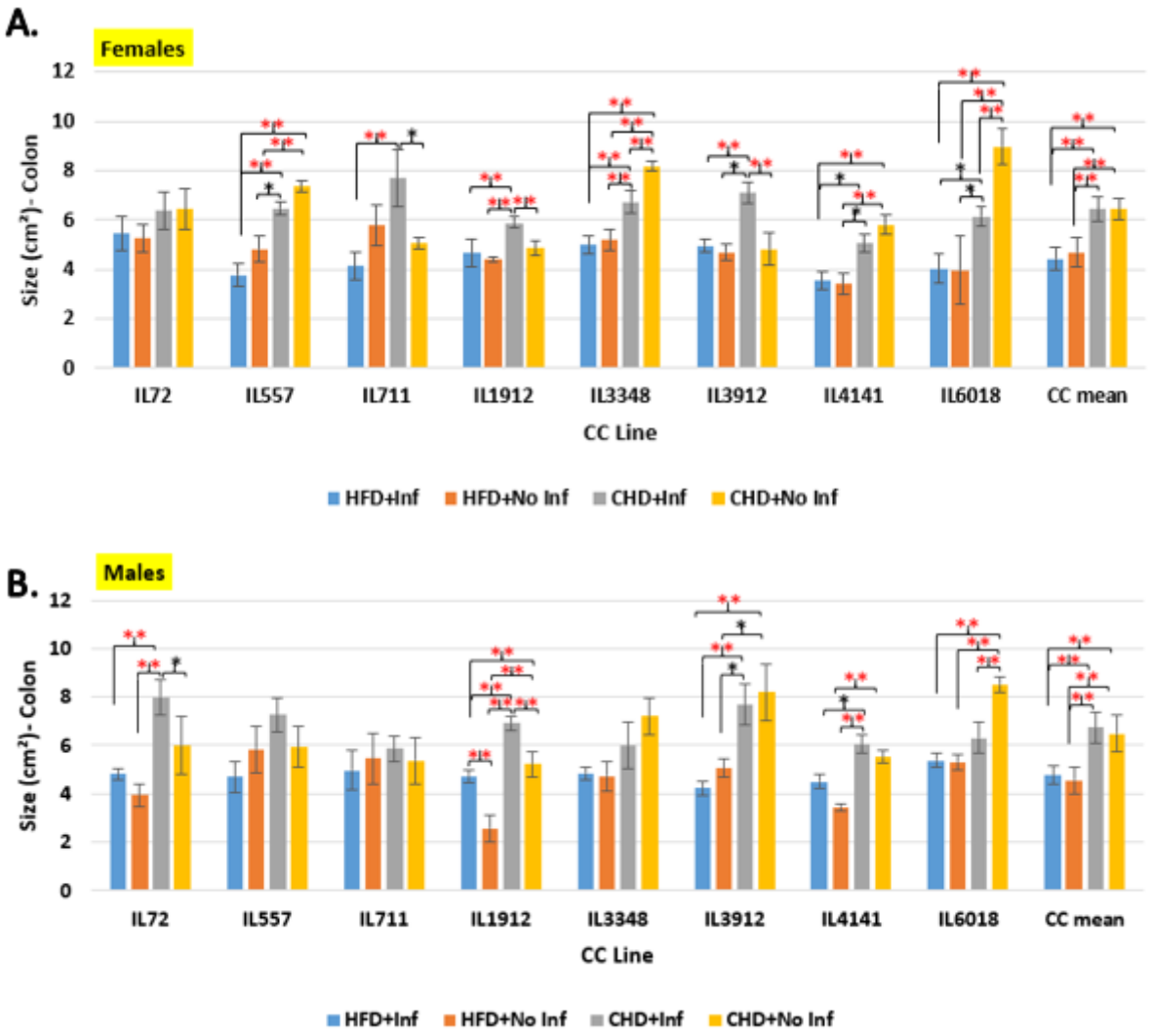


Figure 9

Size of colon in cm² of female and male mice of eight different CC lines and mean population after were maintained for 12 weeks on either chow (18% Fat) diet or high-fat (42% Fat) diet (HFD) and with or without oral infection. X-axis presents CC lines and mean of the entire population; Y-axis presents values of size of the colon in female [A] and male [B] mice.

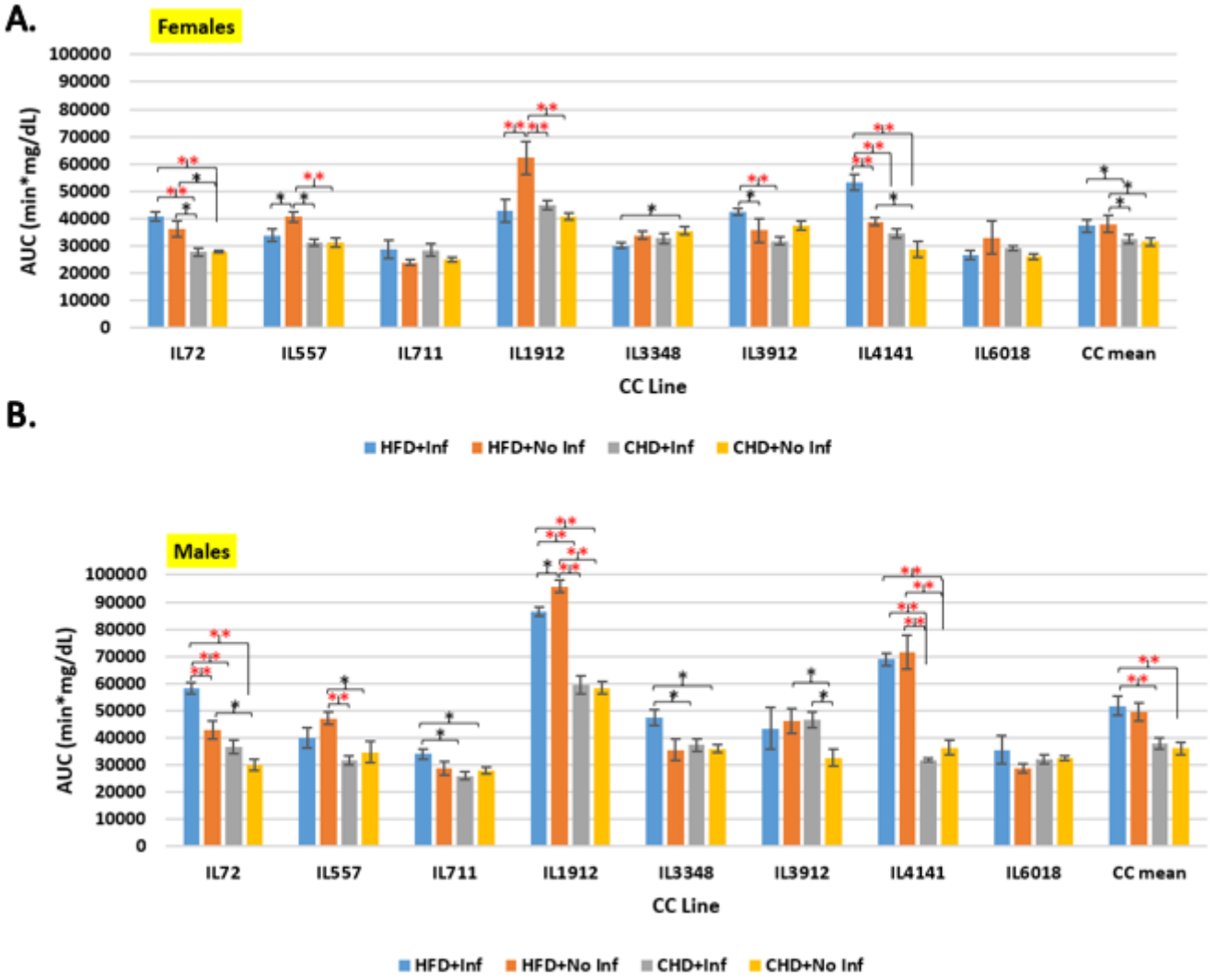
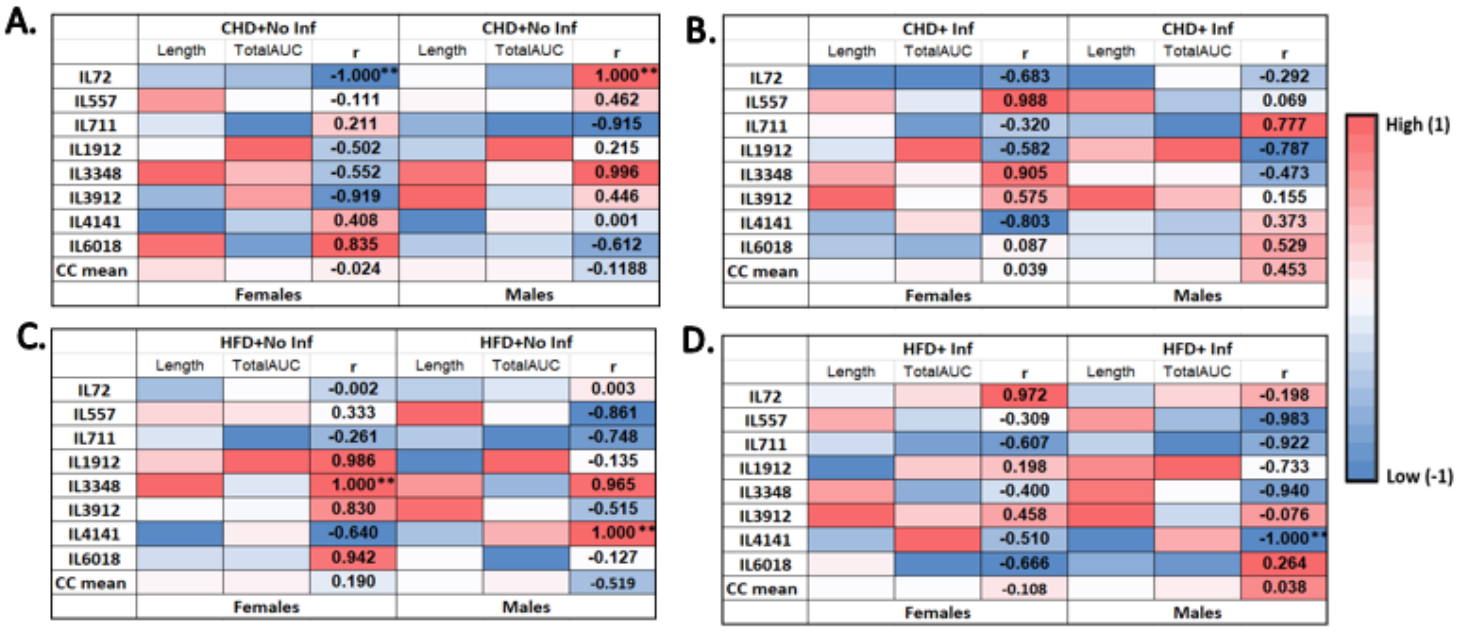


Figure 10

Total area under curve (AUC_{0-180}) of glucose clearance (min*mg/dL) of intraperitoneal glucose tolerance test (IPGTT) of eight different CC lines after maintained either on HFD (42 % Fat) challenge or CHD (18% fat) and with or without oral co-infection. P values less than 0.05 are summarized with one asterisk while P values less than 0.01 summarized with two asterisks. (* $P < 0.05$ and ** $P<0.01$). X-axis presents CC lines and the mean of the entire population; Y-axis presents AUC in female [A] and male [B] mice.



*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).

Figure 11

Heatmap and Pearson correlation between total length of intestines (small and large intestines) and AUC changes of female and male mice of eight different assessed CC lines and the mean of the entire population that were maintained for 12 weeks either on HFD or CHD and with or without oral infection. Figures (A) shows results of non- infected female and male mice maintained on CHD, (B) infected female and male mice maintained on CHD, (C) non infected female and male mice maintained on HFD and (D) infected female and male mice maintained on HFD. The r values ranged between minimum (-1) and maximum (1) and significant values was at p<0.01 (**) and P<0.05 (*).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables1.docx