Identification of genetic risk factors predisposing to the inflammatory oral disease periodontitis

Inaugural-Dissertation
to obtain the academic degree
Doctor rerum naturalium (Dr. rer. nat.)

submitted to
the Department of Biology, Chemistry and Pharmacy
of Freie Universität Berlin

by

Matthias Munz

2018

This work was carried out from February 2015 to September 2018 at the research group Molecular Genetics of Oral Diseases at the Charité - University Medicine Berlin under the supervision of Prof. Dr. Arne Schäfer and at the Institute for Cardiogenetics at the Universität zu Lübeck under the supervision of Prof. Dr. Jeanette Erdmann.

First Referee

Prof. Dr. Arne Schäfer
Department of Periodontology and Synoptic Dentistry
Institute for Dental and Craniofacial Sciences
Charité – University Medicine Berlin
Aßmannshauser Straße 4-6
14197 Berlin, Germany

Second Referee

Prof. Dr. Silke Rickert-Sperling
Cardiovascular Genetics
Charité – University Medicine Berlin
Lindenberger Weg 80
13125 Berlin, Germany

Member of the Department of Biology, Chemistry, and Pharmacy
Freie Universität Berlin

Date of oral examination: Monday, May 13, 2019

Acknowledgments

I owe my deepest gratitude to my supervisor **Prof. Dr. Arne Schäfer**, leader of the research group *Molecular Genetics of Oral Diseases* at the *Charité - University Medicine Berlin*, for introducing me to the interesting world of the complex oral disease periodontitis. Without his continuous optimism, enthusiasm, encouragement and support, this work would hardly have been completed.

I am very grateful to **Prof. Dr. Jeanette Erdmann**, head of the *Institute for Cardiogenetics* at the *University of Lübeck*, for giving me the opportunity to work in her bioinformatics group. It was an honor for me to learn from one of the top human geneticists in Germany.

I am indebted to the whole **ICG team** for the pleasant working atmosphere. All of you supported me in many different ways. I am grateful to my office mates **Dr. Loretto Muñoz**, **Dr. Ilyas Ahmad**, **Dr. Stephanie Tennstedt**, **Dr. Anita Bhandari**, **Syed Haider**, **Martin Gross** and **Viviana Vedder** for the nice discussions, their support and encouragement. I would also like to thank **Dr. Inken Wohlers**, **Dr. Fabian Kilpert**, **Dr. Shengjun Hong** of the research group *Lübeck Interdisciplinary Platform for Genome Analytics* for their advice and discussions during lunch time.

For biostatistical advice at the beginning of my doctorate I wish to present my special thanks to **Dr. Christina Willenborg**.

This dissertation would not have been possible without technical and administrative support. Hence, I would like to thank the system administrators **Dirk Laggin** of the *Institute for Neuro- and Bioinformatics* as well as **Frank Rühlemann** and **Gudrun Amedick** of the *Lübeck OMICS Cluster*. I would also like to thank **Marina Ladewig** and **Dr. Nadine Clausen** of the ICG for their support in organizational and formal matters.

I am very grateful to **Prof. Dr. Silke Rickert-Sperling** for making it possible to do my doctorate at the *Freie Universität Berlin*.

Last but not least, I would like to thank my family, **Ann-Kathrin** and **Oskar**, for understanding my long working hours, for motivating and believing in me.

Sincere thanks to all of you!

Abstract

Background: Periodontitis (PD) is a bacterially induced disease of the oral cavity and, in the long term, can lead to tooth loss due to the inflammatory degradation of the alveolar bone. For the severe forms of PD, there is a prevalence rate of 11% worldwide, making it one of the most common inflammatory diseases. Aggressive periodontitis (AgP) is an early (< 35 years) and particularly severe clinical form of PD, whereas chronic periodontitis (CP) usually has a slow and more moderate course. The estimated heritability of PD is 50%. Twin studies and examples where individuals share the same environment and lifestyle yet have different risk for disease, illustrate the strong influence of hereditary factors on the pathogenesis of PD. Several randomized clinical trials have demonstrated a relationship between PD and coronary artery disease (CAD), with an epidemiological relationship between the two disease to be independent of smoking as a common risk factor. CAD is a complex disease of the arteries with a strong inflammatory component. Plaque formation in the arteries can impair the blood supply, which can trigger a heart attack.

***Status quo*:** Previous studies have identified genetic risk variants at several loci in the human genome that predispose to PD. The gene locus *GLT6D1* is associated with AgP at genome-wide significance. Furthermore, common genetic risk variants of PD and CAD were identified for the gene sites *ANRIL*, *Plasminogen* and *VAMP3*. However, the known genetic risk factors explain only a small fraction of the heritability of PD. Indeed the underlying genetic risk factors remain largely unknown.

Methods & Results: The identification of yet unknown genetic risk factors can significantly contribute to a better understanding of the causes and molecular mechanisms of PD as well as the genetic and molecular relationship to CAD and to the improvement of diagnosis and treatment of PD. In my thesis, I carried out genome-wide association studies and meta-analyses with the aim of systematically identifying further risk genes. I had access to the, to date, world's largest AgP sample with 851 cases and 6,580 controls of Dutch and German descents. In addition, I used a non-genotyped AgP sample with 220 cases and 560 controls of Turkish descent which were used to confirm single genetic variants. For CP samples I had a total of 4,244 cases and 3,328 controls of German and European-American descent which were also taken into account. I could identify new risk variants of PD that either reached genome-wide significance levels ($P < 5 \times 10^{-8}$) or could be independently validated. These variants are located proximal to the genes *DEFA1A3*, *SIGLEC5*, *LOC107964137*, *MTND1P5*, *VAMP8*, *PF4/PPBP/CXCL5*, *IL37*. *In silico* func-

tional characterization of these loci points to specific processes that play a role in wound healing and in bacterial immune defense. A pleiotropic variant at *VAMP8* also increases susceptibility to CAD and suggests a possible molecular mechanism underlying the epidemiological link between PD and CAD. The results confirm that the relationship between PD and CAD cannot be explained solely by shared lifestyle factors. However, the SNP-gene relationships shown in this work require further experimental studies.

Zusammenfassung

Hintergrund: Parodontitis (PD) ist eine bakteriell induzierte Erkrankung der Mundhöhle und kann durch den entzündungsbedingten Abbau des Alveolarknochens langfristig zum Zahnverlust führen. PD zählt mit einer Prävalenzrate von 11% weltweit für die schweren Formen zu den häufigsten Entzündungskrankheiten. Aggressive PD (AgP) ist ein sehr früh (< 35 Jahre) auftretendes und besonders schwer verlaufendes Krankheitsbild der PD, chronische Parodontitis (CP) hingegen hat einen langsamen moderateren Verlauf. PD hat eine geschätzte Erblichkeit von 50%. Zwillingsstudien und Beispiele von Personen, die innerhalb derselben Umwelt mit einem gleichen Lebensstil leben, aber ein sehr unterschiedliches Krankheitsrisiko haben, verdeutlichen den starken Einfluss erblicher Faktoren auf die Pathogenese der PD. Ein Zusammenhang zwischen dem Vorliegen einer PD und einer koronaren Herzkrankheit (KHK) konnte in mehreren randomisierten klinischen Studien nachgewiesen werden und zeigt, dass der epidemiologische Zusammenhang zwischen beiden Erkrankungen unabhängig vom gemeinsamen Risikofaktor Rauchen ist. Die KHK ist eine komplexe Erkrankung der Arterien, mit einer starken Entzündungskomponente. Durch Plaquebildung in den Arterien kann die Blutversorgung beeinträchtigt werden, wodurch ein Herzinfarkt ausgelöst werden kann.

***Status quo*:** In vorangegangenen Studien konnten bereits Einzelnukleotid-Polymorphismen (SNPs) an mehreren Genorten im humanen Genom identifiziert werden, die für die PD prädisponieren. Der Gen-Lokus *GLT6D1* ist mit genomweiter Signifikanz mit AgP assoziert. Außerdem wurden gemeinsame genetische Risikovarianten der PD und der KHK für die Genorte *ANRIL*, *Plasminogen* und *VAMP3* identifziert. Die bekannten genetischen Risiko-Loci erklären jedoch nur einen kleinen Teil der Heritabilität der PD. Tatsächlich sind die der PD zugrunde liegenden genetischen Risikofaktoren jedoch noch weitgehend unbekannt.

Vorgehen & Ergebnisse: Die Identifikation noch unbekannter genetischer Risikofaktoren kann wesentlich dazu beitragen, die Ursachen und molekularen Mechanismen der PD sowie den genetischen und molekularen Zusammenhange zwischen PD und KHK besser zu verstehen und die Diagnose und Therapie von PD zu verbessern. In der vorliegenden Arbeit wurden genomweite Assoziationsstudien und Meta-Analysen durchgeführt, mit dem Ziel, systematisch weitere Risikogene zu identifizieren. Dazu wurde das aktuell weltweit größte AgP-Sample mit 851 Fällen und 6.580 Kontrollen mit niederländischen und deutschen Teilnehmern

verwendet. Zudem wurde ein für die Bestätigung einzelner genetischer Varianten ein nicht-genotypisiertes AgP-Sample mit 220 Fällen und 560 Kontrollen türkischer Herkunft hinzugenommen. Mit insgesamt 4.244 Fällen und 3.328 Kontrollen standen außerdem CP-Samples deutschen und europäisch-amerikanischen Ursprungs zur Verfügung. Es konnten neue Risikovarianten der PD identifiziert werden, die genomweites Signifkanzniveau ($P < 5 \times 10^{-8}$) erreichen oder unabhänigig validiert werden. Diese Varianten befinden sich proximal zu den Genen *DEFA1A3*, *SIGLEC5*, *LOC107964137*, *MTND1P5*, *VAMP8*, *PF4/PPBP/CXCL5*, *IL37*. Die funktionelle *In Silico*-Charakterisierung dieser Loci weist auf spezifische Prozesse hin, die in der Wundheilung sowie in der bakteriellen Immunabwehr eine Rolle spielen. Eine pleiotrope Variante bei *VAMP8* erhöht außerdem die Anfälligkeit für KHK und weist auf einen möglichen molekularen Mechanismus hin, der dem epidemiologische Zusammenhang der PD und KHK zugrunde liegen kann. Die Ergebnisse bestätigen, dass der Zusammenhang von PD und KHK nicht ausschließlich durch geteilte Lifestyle-Faktoren zu erklären ist. Die in dieser Arbeit aufgezeigten SNP-Gen-Beziehungen erfordern jedoch weitergehende experimentelle Studien.

Contents

List of Figures

Abbreviations

AGE	Advanced Glycation End Product
AgP	Agressive Periodontitis
CI	Confidence Interval
CP	Chronic Periodontitis
CAD	Coronary Artery Disease
DEG	Differentially Expressed Gene
DNA	Deoxyribonucleic Acid
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
ETL	Extraction, Transformation, Loading
eQTL	Expression Quantitative Trait Loci
GWAS	Genome-Wide Association Study
LD	Linkage Disequilibrium
OR	Odds Ratio
PD	Periodontitis
QC	Quality Control
QTL	Quantitative Trait Loci
RAGE	Receptor for Advanced Glycation End Products
SNP	Single Nucleotide Polymorphism
WHO	World Health Organization

Chapter 1

Introduction

The oral cavity forms the entrance to the gastrointestinal tract and offers excellent conditions for microorganisms due to the high humidity, the constant temperature and the continuous food supply. The food is seized, crushed and ground by the teeth. The area that joins teeth and gingival tissue is particularly susceptible to pathogens and a particular challenge for the body's immune system. This is also illustrated by the coinciding development of adaptive immunity and teeth in early jawed vertebrates during evolution [1, Chapter 15] [2, 3]. The healthy oral cavity is inhabited by a very complex and species-rich community. It is assumed that for oral inflammatory diseases these pathogens reflect a dysbiosis of the microflora, i.e. by a state of disequilibrium in the community of oral microbes (reviewed in [4]). As a result, the immune system triggers an inflammatory reaction which can become chronic if no homeostasis occurs. As a consequence, the inflammatory substances can extensively and irreversibly damage the patient's supporting tissue of the teeth, the periodontium, a pathological process that is known as the oral disease periodontitis (PD) (**Figure 1.1**).

PD is among the common inflammatory diseases, with an estimated heritability of 50% [5]. It affects human populations with a worldwide prevalence rate of 11% in adults with aged 30 years and older for the severe forms [6] and is the major cause of tooth loss in adults above 40 years [7]. Since PD is painless, the inflammation usually persists long-term making it common for PD to have reached advanced degrees of severity before it's diagnosed and eventually treated (**Figure 1.2**). Since the inflammation can affect large areas of the oral mucosa, it can pose a substantial strain for the immune system. Apart from the personal implications, PD also can be a substantial burden of public health care [8].

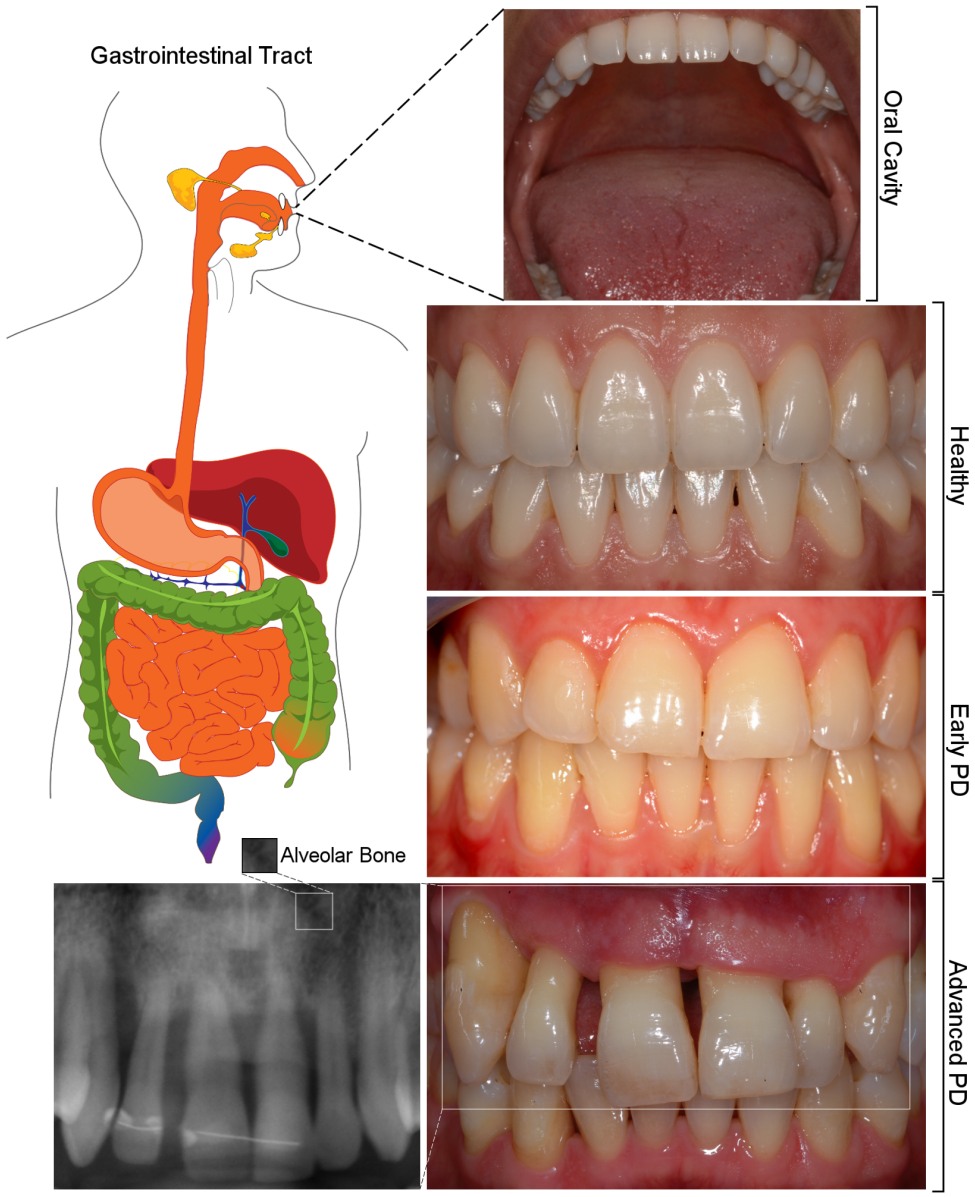

Figure 1.1: Healthy and diseased periodontium. Periodontitis is a common inflammatory disease of the supporting tissue of the teeth in the oral cavity. The oral cavity, equipped with special set of tools such as teeth and salivary glands, represents the first step in the digestive system. It is a warm and unsterile space offering a perfect conditions for microbes which, in a healthy cavity, form a well-balanced and necessary microflora environment. A dybiosis of the microflora, i.e. through external factors like smoking or oral hygiene, can trigger an inflammation of the gingiva that can lead to an early PD (Early PD). If equilibrium of the microflora is no longer restored, it may lead to the irreversible degradation of the supporting tissue of the teeth including resorption of alveolar bone (Advanced PD). The dental images were provided by the Department of Periodontology and Synoptic Dentistry at the Charité - University Berlin.

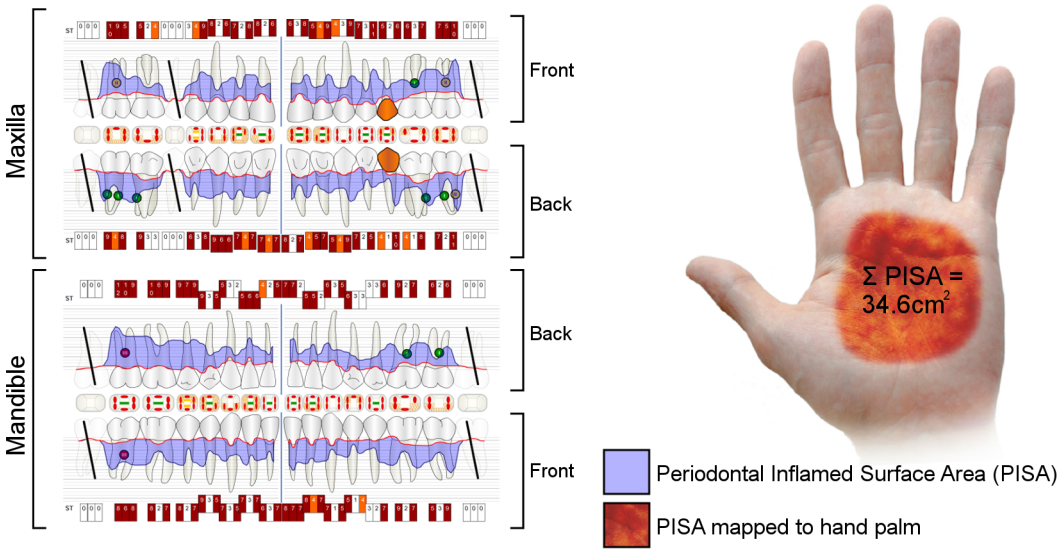

Figure 1.2: Inflammatory burden of periodontitis patients. In periodontitis, the inflammation of the oral mucosa can reach a large extent and be a strong personal burden for the patients. (Left) The oral inflamed area of a female PD patient with 50 years of age corresponds to $34.6cm^2$, covering 90% of the total epithelial surface ($38.3cm^2$). The inflamed area was estimated using the periodontal inflamed surface area (PISA) method which takes the clinical parameters clinical attachment level, recessions and bleeding on probing into account [9]. (Right) Putting the entire inflammable area together, it covers large parts of an adult male hand palm. The dental images were provided by the Department of Periodontology and Synoptic Dentistry at the Charité - University Berlin.

Long-term smoking, excess dietary fat and diabetes contribute to the microbial shift of the oral microbiota, which is considered to trigger the inflammatory reaction, and the overall risk of PD. The inflammation leads to gingival bleeding, gingival pocket formation, and to the resorption of alveolar bone which withdraws from the site of inflammation, a process that can lead to tooth loss. These steps are unique in the pathogenesis of complex diseases, because the oral cavity is the only part of the human organism where, in an intricate microbial environment of microbes, soft tissues such as the gingival mucosa dynamically interact with hard tissues like teeth and skeletal bone.

However, the precise molecular mechanisms that drive the individual steps in the pathogenesis of PD are currently unknown. Moreover, unraveling the genetic architecture of PD may help to understand the molecular mechanisms that underlie the disease risk.

1.1 Manifestations of periodontitis

PD is classified into the widespread form of chronic periodontitis (CP) and the rare and much more severe form of aggressive periodontitis (AgP). AgP has a prevalence of <0.1% globally [10]. CP primarily affects adults and its prevalence increases with age [11, 12], whereas AgP is characterized by an early age of onset, affecting mainly adolescents or young adults [13, 14]. CP and AgP have a similar aetiology and histopathology and can be considered as parts of the same disease spectrum (**Figure 1.3**). AgP has a stronger and better-established heritable component [15–17]. It is considered that genetic factors play a stronger role in disease onset and progression of AgP, because of the usually very young age of onset, implying the absence of other risk factors like long-term smoking or diabetes, and often concomitant with the absence of plaque.

Figure 1.3: Manifestations of periodontitis. Periodontitis can be classified into the early-onset and very severe form of aggressive periodontitis, and the late-onset and less severe form of chronic periodontitis. Generally, early-onset forms of complex diseases are thought to be determined by genetic factors to a larger extent.

The extent and severity of PD is determined by measuring different parameters of the periodontium [18]. The probing pocket depth of the gap between the gums and teeth is measured by sliding a dental probe under the gum line next to the teeth (**Figure 1.4**). The clinical attachment loss tells, how strong the teeth are attached to periodontium by periodontal ligament fibers. It is derived by measuring the distance from the cemento-enamel junction to the base of the pocket. Since in PD the inflammation can occur in a repetitive manner with periods of reduced inflammation, the pocket depth can vary and influence both probing depth and clinical attachment loss. The loss of alveolar bone is measured using dental X-rays. Finally, the overall number of affected teeth are also a crucial parameter.

However, no consistent thresholds for parameters describing the state of the periodontium have been used so far. Accordingly, no universally accepted definition of periodontitis exists. Also, no disease-specific biomarkers have been discovered. Consequently, the differentiation between chronic and aggressive periodontitis remains

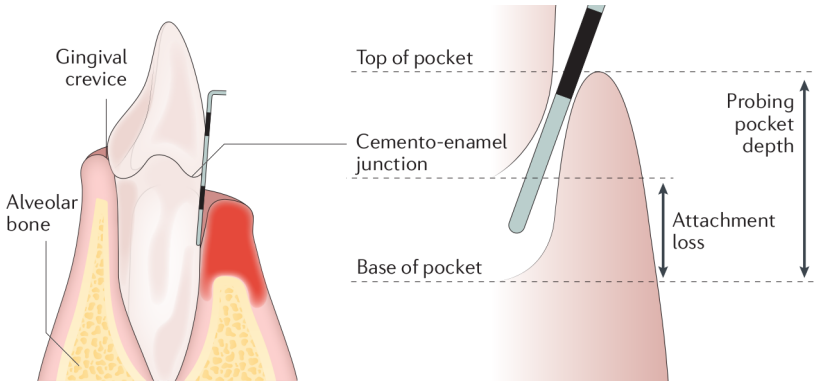

Figure 1.4: Probing pocket depth. (Left) Schematics of a tooth with healthy gingiva on the left and advanced periodontitis on the right side which is indicated by secluded gingiva and loss of alvolar bone [18]. (Right) Gap between gums and tooth in which a steel narrow-diameter probe is passed to measure the pocket depth.

unclear.

Recently in 2018, a new classification scheme for PD has been released which will replace the definitions of AgP and CP by a multidimensional staging and grading system [19] (reviewed in [20]). Since the new classification scheme was released after I conducted the analyses in this thesis, it is not further considered in the following.

1.2 Shared epidemiology of PD and other diseases

Recurrent and persistent inflammations caused by bacteria are also recognized as continually renewing reservoirs for the systemic spread of bacterial antigens, cytokines, and other pro-inflammatory mediators and may bring a burden onto the rest of the body [21]. Accordingly, researchers have hypothesized the etiologic role of PD in the pathogenesis of various systemic illnesses like diabetes mellitus (reviewed in [22]), osteoporosis (reviewed in [23]) and cardiovascular disease (CVD) [24], bridging the once-wide gap between medicine and dentistry.

Diabetes mellitus Diabetes mellitus is a complex metabolic disease characterized by hyperglycaemia due to defects in insulin metabolism. The World Health Organization (WHO)[1] ranks diabetes as the seventh leading cause of death worldwide in 2013. There is strong evidence of a bi-directional causation between diabetes and PD. It has been shown that inflammation promotes the development of insulin resistance by pro-inflammatory mediators like TNFα, IL6, IL1 and interferons [25, 26]. A mouse model of diet-induced obesity, in which the inhibition of TNFα

[1] http://www.who.int

prevents the onset of obesity-associated insulin resistance, illustrates the effect of pro-inflammatory factors on the pathogenesis of diabetes [27]. Pro-inflammatory mediators have increased expression levels in periodontal inflammation [28–30]. Conversely, the finding that periodontitis in patients with diabetes is not only more frequent but also more severe than in healthy people, indicates a causative role of diabates on PD (reviewed in [31, 32]). Several hypthoses exist describing this role, however they remain controversial, since none have been confirmed. A mechanism that proposes a direct causality of diabetes on periodontitis may act through the effects of advanced glycation end products (AGEs) which are formed to an accelerated degree in diabetic tissues due to chronic hyperglycemia. AGEs can bind to special plasma membrane localized receptors for AGEs (RAGEs), which are expressed on different cells, e.g. on macrophages. Macrophages are then transformed into hyperreactive cells that produce pro-inflammatory cytokines, which can drive hyperinflammatory responses, vascular modifications and altered healing, contributing to the development of periodontitis [33]. To-date, there are 73 known genetic risk loci that reach with genome-wide significance ($P < 5 \times 10^{-8}$) for type 2 diabetes [34].

Osteoporosis Osteoporosis is a systemic skeletal disorder that is characterized by a decrease in bone density as a result of the degradation of bone tissue in excess of the build-up during natural bone remodeling increasing the risk of fractures [35]. Osteoporosis has prevalence of 10% in the US adult population aged 50 years and older [36].

Osteoporis shows commonalities with PD (reviewed in [23]). For both diseases, excessive bone resorption is a central feature. Moreover, PD and osteoporosis share the risk factor smoking. However, a causal relationship between osteoporosis and PD has not been established yet.

Coronary artery disease According to the WHO, CVD is the most frequent cause of death globally. CVD describes a class of disorders of the heart and blood vessels, with coronary artery disease (CAD) being the most common. CAD is the manifestation of atherosclerosis in the coronary arteries supplying the heart muscle, also known as myocardium, with oxygen and nutrients (reviewed in [37]). Its pathogenesis comprises of chronic inflammation of the arterial wall caused by an unbalanced lipid metabolism and a maladjusted immune response. Due to the disturbed balance of lipid accumulation, immune reactions and their release, plaque forms, which leads to a thickening of the arterial wall. This may result into reduced

blood flow inside the arteries affecting the supply and eventually lead to myocardial infarction. Thus, analoguous to PD, CAD has a strong inflammatory component (reviewed in [38]). An association between CAD and PD was reported in several clinical and observational studies (reviewed in [39]). Because of the high prevalence of CAD and PD and the high mortality of CAD, this association is potentially of public health importance. Recent evidence indicates that the observed association between CAD and PD is independent of smoking [39] and obesity [40]. However, it could be explained in part by other shared risk factors like diabetes and age. Unlike diabetes and PD, little is known about the presence of a causal relationship of CAD and PD. One indication that PD may play a direct causal role in CAD is the discovery of oral commensial bacteria in artheroslerotic plaque of CAD patients [41] (reviewed in [42]). It is believed that this is the result of an arterial infection by oral bacteria that were able to enter the bloodstream in large quantities due to inflammation of the periodontium. However, the exact mechanisms underpinning this relationship remains unclear. In this context, the knowledge of shared genetic risk variants could substantially contribute to the understanding of the mechanisms that underlie the epidemiological associations.

Due to the high mortality of CAD and the demographic development especially in western countries, there is a special research interest in this disease. For this reason, CAD already has a well-established genetic basis. Large consortia such as the "Coronary ARtery DIsease Genome wide Replication and Meta-analysis plus The Coronary Artery Disease Genetics"[2] (CARDIoGRAMplusC4D) have been founded, pooling patient cohorts into large meta-analyses. With the help of such large sample sizes, 163 genetic risk variants of genome-wide significance ($P < 5 \times 10^{-8}$) for CAD could be discovered so far (reviewed in [43]), including risk variants with very low effect sizes (i.e. Odds Radio [OR] < 1.05, **Figure 1.5**).

1.3 Known genetic risk factors of PD

Earlier studies identified a single nucleotide polymorphism (SNP) within the gene *GLT6D1* to be associated with AgP at a genome-wide significance level [44], which was later replicated in a sub-Saharan population from Sudan [45].

Further analyses demonstrated a shared risk locus of AgP and CAD at *PLG* (plasminogen) [46, 47] and a shared risk locus of AgP and CAD at *ANRIL* (*CDKN2B-AS1*) [48–52]. In addition, evidence for a shared association of a rare genetic variant

[2] http://www.cardiogramplusc4d.org/

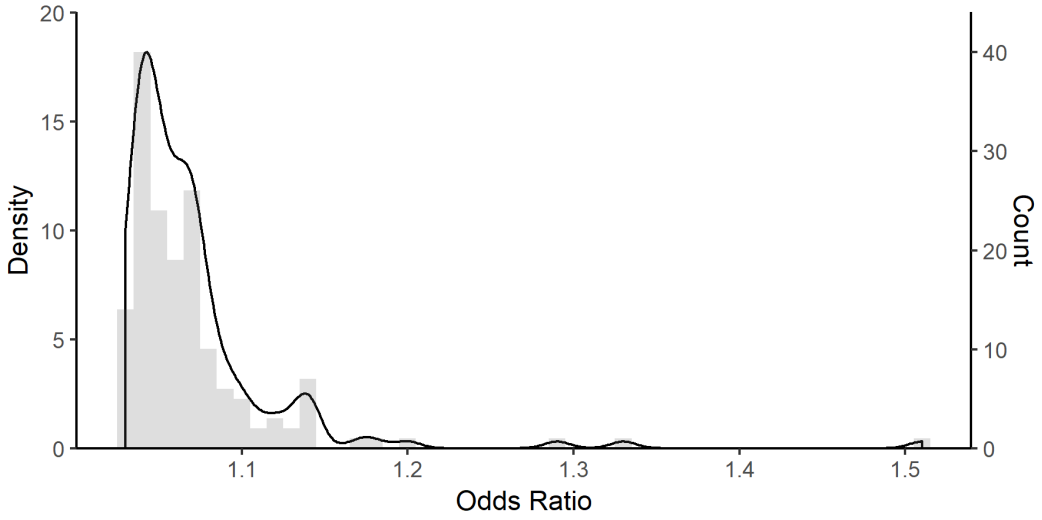

Figure 1.5: Frequency distribution of the effect sizes (odds ratio) of 163 known CAD risk variants. Number of variants increases with lower effect sizes. Minimum: 1.03; 1st quartile: 1.04; mean: 1.07; 3rd quartile: 1.073; maximum: 1.51; Variant information was taken from [43].

at *VAMP3* with AgP and CAD was observed [53]. *VAMP3* is located in a chromosomal region that was earlier described to be associated with increased colonization of oral periodontal pathogens [54].

The *PLG* locus was identified as shared risk variant of AgP and CP [55]. GWAS that focused on CP only, suggested various risk variants but have failed to give clear statistical evidence for association [16, 54, 56–63]. This may be partly due to the high heterogeneity of CP but probably mostly relates to the limited statistical power of these GWAS.

The evidence of a systemic link between PD and both diabetes and osteoporosis has not yet been established at the genetic level so far.

1.4 General methods

1.4.1 Genome-wide association studies

A genome-wide association study is a research study in which a genome-wide set of small genetic variants, mainly SNPs, is systematically investigated for associations with a phenotype in a sample of unrelated individuals (reviewed in [64–66]). This kind of study became popular in the mid-2000s, when advances in genotyping array technologies allowed to type hundreds of thousands of human SNPs in a single

experiment and at comparably low costs.

The large number of SNPs detected by genotyping arrays still represents only a fraction of all the variation that is caused by SNPs in the human genome. However, the fact that many SNPs are linked with each other (also known as linkage disequilibrium [LD], described in more detail in chapter 1.4.4 on page 14) in human populations allows to explain a large amount of common genetic variation. For example, 500,000 SNPs are required to completely capture common variation (i.e. minor allele frequency [MAF] $> 5\%$) in Utah Residents with Northern and Western European ancestry (CEU) [67].

The amount of SNPs that are tested in a GWAS increases the number of false-positives and thus the false-positive (type 1) error α. Therefore, adjusting for multiple testing plays a crucial role in GWAS and usually, Bonferroni correction is applied to control the family-wise error rate. This is done by adjusting the significance level α to α/n where n represents the number of tests. Putative association signals can be confirmed by replicating it in an independent study [68]. Then, the significance level of the replication has to be adjusted by the number of replicated variants. Alternatively, a meta-analysis (described in subsection 1.4.2 on page 12) of two or more independent studies and the application of the genome-wide significance threshold to the pooled P-value is also considered as validation method [69, 70]. A P-value below 5×10^{-8} is generally accepted as being genome-wide significant [70, 71]. In terms of Bonferroni correction, this threshold corresponds to 1 million independent SNPs given a unadjusted significance level of $\alpha = 0.05$ [72].

Depending on whether the phenotype is binary such as disease phenotypes like CAD or PD or quantitative such as body height, logistic or generalized linear regression models can be used to model the relation between a genetic variant and the phenotype. A simple linear regression model of a SNP X_i and a phenotype Y_i can be formulated as $Y_i = \beta_0 + \beta_1 X_i$ where $i = 1...n$ and n represents the number of samples. In other words, a straight line is placed through the data points that best describes the relationship between the variables X and Y. In this two dimensional example, β_0 represents the intercept on the y axis and β_1 the slope of the line. The beta coefficient of a SNP represents its effect size and the signs provide information on the direction of action. When the phenotype is binary, the response variable Y are either 0 or 1. The prediction of Y can be seen as probabilities in $[0, 1]$: $Y_i = P(Y_i = 1) = \beta_0 + \beta_1 X_i$. To estimate the probability for a linear combination of independent variables, the linear function has to be adapted by applying the *logit* function onto Y: $logit(P(Y_i = 1)) = ln(P(Y_i = 1)/(1 - P(Y_i = 1))) = \beta_0 + \beta_1 X_i$.

The *logit* is the natural logarithm of the chance ($Odds = P(Y_i = 1)/(1 - P(Y_i = 1))$ that $Y_i = 1$ and equivalent to a linear function of the independent variables. In logistic regression, there is a direct relationship between beta coefficients and odds ratios, which allows to transform beta coefficients into odds ratios ($OR = exp(\beta)$) and vice versa ($\beta = ln(OR)$). For both model types, a P-value for the association level of a SNP can be computed from the beta estimate $\hat{\beta}$, which corresponds to β_1 in the previous example, and its standard error $SE(\hat{\beta})$ by using the Wald statistic, i.e. by calculating $2 \times \Phi(-|\hat{\beta}|/SE(\hat{\beta}))$, where Φ denotes the distribution function for the normal distribution.

Single genetic variants such as SNPs can be modeled in different ways. In the additive model which is the standard model, the number of alternative alleles (A) is counted for each SNP. In a diploid genome, this results either in 2 (AA), 1 (AB) or 0 (BB) with B stating the reference allele. Other genetic models are recessive (1 [AA], 0 [AB], 0 [BB]), dominant (1 [AA], 1 [AB], 0 [AB]).

The statistical power or sensitivity of a GWAS study, i.e. the probability of correctly rejecting the null hypothesis (H_0: no association between SNP and phenotype), depends on the false-negative (type 2) error β (power= $1 - \beta$), the sample size, the genetic model, risk allele frequencies and effect sizes. Since the characteristics about the causal variants are unknown prior to a GWAS, minimum thresholds for frequencies and effect sizes are calculated such that SNPs with these attributes can be detected with reasonable power. For one-stage GWAS, the statistical power for case-control studies can be calculated using the Genetic Association Study (GAS) Power Calculator[3]. Estimates of genetic effects of newly found associations tend to be upwardly biased owed to a phenomenon which is known as the "winner's curse" [73]. The "winner's curse" describes a phenomenon in which winners in competing auctions tend to pay more for an item than its actual value. In relation to GWAS, this means that genetic variants with an effect size for which the study is underpowered, it can only be detected, if the real effect size is overestimated. Overestimated effect sizes can result in replication sample sizes being selected too small and in a replication failure, i.e. producing a false negative result. The "winner's curse" can also occur in GWAS meta-analyses, when the inter-sample heterogeneity is too high.

Genotypes require extensive quality control (QC), before a GWAS analysis can be performed. QC steps include updating of SNP identifiers and their genetic position, allele strand flipping and reference/alternative allele swapping, all based on a preselected reference genome build. Moreover, a sex check based on the genotypes can

[3] http://csg.sph.umich.edu/abecasis/gas_power_calculator/index.html

help to identify possible phenotyping errors or mistaken identities. Outliers in the population structure can be identified and removed using principal component analyses (PCR) or multidimensional scaling (MDS). Identity-by-descent and identity-by-state analyses allow for detecting cryptic relatedness. Additionaly, several filters (SNP callrate, sample callrate, minor allele frequency, Hardy-Weinberg-equilibrium, heterozygosity) can be applied to further improve the quality of the variant and sample set. PLINK 1.90[4] is a software program offering a wide range of tools to perform QC on genotypes [74].

Association estimates can be biased by unknown and known confounding factors like population admixture, gender or lifestyle factor such as smoking for PD. The amount of population admixture is measured by PCR or MDS. The effect of these factors on GWAS results is assessed by the genomic inflation factor (λ) [75]. Lambda is defined as the ratio of the median observed test statistic across all tested SNPs to the expected median test statistic. The expected median test statistic for a chi-squared (χ^2) test with one degree of freedom is approximately 0.456. A $\lambda > 1$ indicates a bias through confounding factors. Then, genomic control can be applied to adjust the test statistics such that $\lambda_{adjusted} = 1$ by dividing each test statistic with the original λ [76]. Known confounding factors can also be considered in the regression model by adding as additional independent variable beside the genetic variable.

Due the beforementioned large extent of LD in human populations, identified associations often cannot be broken down to a single causal SNP but to a haplotype block harboring multiple highly correlated/linked SNPs. Moreover, approximately 90% of the identified SNPs are located in noncoding DNA regions, further analyses are required to identify the specific casual genes [77]. The prioritization of genes by means of published biological datasets (see subsection 1.4.4 on page 14) is used to illuminate causality.

GWAS have been very successfull in identifying associations of common variants with complex diseases and traits and helped substantially to understand underlying (patho-)mechanisms [78]. As an example of it's relevance in translational processes, it has been shown that the proportion of drug mechanisms with drug targets that are supported by genetic associations, significantly increases along the drug development pipeline and the chances for drug approval [79].

[4] https://www.cog-genomics.org/plink2

1.4.2 Meta-analysis of GWAS results

The results of multiple independent GWAS on the same phenotype can be combined in a meta-analysis to increase statistical power and reduce false-positive findings (reviewed in [68, 80]). Meta-analyses can use the SNP-specific summary statistics and do not rely on individual-level genotype data.

The most powerful meta-analysis model for detecting associations is the fixed effects model (FEM). It assumes that the true effect size of an association is the same in all studies, implying a low inter-study heterogeneity. Therefore, this model is not appropriate anymore, if the heterogeneity is too high. Instead, the far less powerful random effects model (REM) can be used which specifically accounts for heterogeneity.

Several sources of heterogeneity exist such as differing population structures and phenotype definitions across the studies. While different population structures can be handled by adjusting for principal components of the population structure, harmonization of different phenotype definitions is often not possible.

The *Cochran's Q* statistic and the heterogeneity index I^2 can be used to measure the level of heterogeneity (reviewed in [81]). Q is the weighted sum of squared differences between individual study effects and the pooled effect resulting from the meta-analysis. When using an inverse variance study weighting in the models above, the weights represent the inverse variance of each individual study. It follows a X^2 distribution with the number of studies minus 1 degrees of freedom (df) (H_0: no heterogeneity). I^2 is calculated as $I^2 = 100\% \times (Q - df)/Q$ and describes the percentage of variation that can be attributed to heterogeneity. Values exceeding 0.5 indicate a moderate to high heterogeneity.

One important step prior to a meta-analysis is to harmonize the datasets, for example by making sure that all studies are on the same reference build and to have clarity which is the effect allele (allele to which the effect refers) and non-effect allele for each specific dataset.

1.4.3 Genotype imputation

The imputation of genotypes is a technique in which the genotypes, which had not directly been genotyped, are predicted from genotypes that were directly assayed using genotyping arrays (reviewed in [82, 83]). This technique works by finding shared haplotype blocks between the study sample and a reference haplotype panel.

Since the reference contains very densely genotyped samples, it can then be used to derive missing genotypes in the study sample. Pre-phasing, i.e. the separation of estimating the haplotypes of the study sample and deriving missing or untyped genotypes in shared haplotype blocks into two steps, can reduce the computational costs dramatically [84].

Imputation with pre-phasing is the most popular method. Two commonly used software frameworks for this method are MaCH [85] plus Minimac [86] and ShapeIt v2 [87] plus Impute2 [84, 88], whereby the first is for pre-phasing and the second for imputation.

Haplotype references panels are provided by three consortia. The HapMap3 panel provided by *The International HapMap Project*[5] includes 1.6 million common small variants (i.e. SNPs and Indels) in 1,184 individuals from 11 populations [89]. The 1000 Genomes Phase 3 panel by *The 1000 Genomes Project*[6] contains 80 million small variants in 2,504 individuals from 27 populations [90–92]. The third reference panel from the *The Haplotype Reference Consortium*[7] includes 40 million small variants in 38,821 individuals with predominantly European ancestry [93]. It combines genotype data from 20 cohorts including the 1000 Genomes cohort and was created to increase imputation accuracy for rare variants down to a frequency of 0.1%.

Incorporating the predicted *in silico* genotypes together with the typed genotypes in a subsequent GWAS has several advantages. It can increase the number of SNPs to be tested for association and boost association signals at causal loci. Additionally, the larger number of association signals allows to resolve or fine-map causal variants more precisely. Moreover, when combining multiple GWAS of different genotyping arrays in a meta-analysis, genotype imputation can increase the amount of overlapping genotypes and improve the results as described. Also, genotype imputation has the ability to detect and correct for genotyping errors.

Recently, two imputation webservers were released, moving the computationally expensive imputation procedure to the cloud: the European *Sanger Imputation Service*[8] and the *Michigan Imputation Server*[9] which represents the US counterpart [93, 94].

[5] https://www.genome.gov/10001688/international-hapmap-project/
[6] http://www.internationalgenome.org/
[7] http://www.haplotype-reference-consortium.org/
[8] https://imputation.sanger.ac.uk
[9] https://imputationserver.sph.umich.edu

1.4.4 *In silico* annotation of genetic variants

Linkage disequilibrium Linkage equilibrium refers to a random association of alleles at two loci whereas linkage disequilibrium (LD) referes to non-random association of the same [95]. Several factors influence LD, including recombination, mutation, genetic drift and founder effect [96, 97].

Given markers A and B with alleles A1, A2, B1, B2, allele frequencies p_A1, p_A2, p_B1 and p_B2 and pairwise allele frequencies p_{A1B1}, p_{A2B2}, p_{A2B1} and p_{A1B2}, the coefficient for linkage disequilbrium D is calculated as $D = p_{A1B1}p_{A2B2} - p_{A2B1}p_{A1B2}$. The strength of association is often indicated as the standardized disequilbrium D-prime (D') ranges between -1 and $+1$ [98]. $D' = D/D_{max}$ if D is positive, otherwise $D' = D/D_{min}$ ($D_{max} = min(p_{A1}p_{B2}, p_{A2}p_{B1})$; $D_{min} = max(-p_{A1}p_{B1}, -p_{A2}p_{B2})$). Another measure of LD is the correlation coefficient r which is calculated as $r = D/\sqrt{p_{A1}p_{A2}p_{B1}p_{B2}}$. D' provides information on historic recombination events, whereas r^2 provides information on the correlation of two markers and is the measure of interest for GWAS. If $r^2 = 1$, then A and B are in perfect LD and have exactly the same MAF. If $|D'| = 1$, then A and B are said to be in complete LD.

In GWAS, LD is calculated for pairs of SNPs using genotype data. Since LD structure varies between different populations, it is necessary to use LD information based on an appropriate population. LD information can be calculated with PLINK 1.9 [74].

eQTL Expression quantitative trait loci (eQTLs) are genomic regions containing genetic variants which influence the expression of one or more genes (reviewed in [99]). Since most of GWAS associations are common variants in non-coding regions, it is not conclusive which gene is affected by them. EQTLs can help to relate these variants with genes and to reveal the underlying biological mechanisms.

Similar to GWAS analyses, in eQTL analyses SNPs are associated with phenotypes, i.e. the expression of genes. However, in GWAS analyses usually SNP associations with only one phenotype is investigated at the same time, whereas in eQTL analyses the expressions of multiple genes are taken into account. This results in a higher number of statistical tests and requires a stronger adjustment for multiple testing. Supposing an analysis of 1 million SNPs and 20,000 genes, the genome-wide significance level for an eQTL is $2.5 \times 10^{-12} = 0.05/1,000,000/20,000$ when applying Bonferroni correction for a nominal significance level of 0.05, compared to a significance level of 5×10^{-8} for GWAS.

eQTLs are classified according to the relative sites to the genes they influence and according to their mode of action. The mode of action of an eQTL with regard to a target gene are either *cis*, if it affects the expression of the target directly, e.g. by modifying a transcription factor binding site, or *trans*, by affecting another gene that has regulatory effects on the target gene. Since mode of action is usually unknown prior to the functional characterization of an association, the mode of action is usually estimated by the distance between eQTL and target gene. However, the *cis* mode can also be present for more distant eQTLs and similarly, the *trans* mode can also be present for eQTLs in proximity to the target gene.

Just like RNA levels, other processes along the gene expression cascade can be studied in this way as well, including DNA methylation, histone modification, transcription factor binding, active transcription, translation and protein levels.

Several public databases exist for eQTL data: the Genotype-Tissue Expression project (GTEx) Portal [100], Haploreg (Ward2016), GRASP [101], GEUVADIS [102], SCAN [103], seeQTL [104], Blood eQTL Browser [105], ExSNP [106] and BRAINEAC [107]. Moreover, the database pGWAS [108] contains QTL data on protein abundance (pQTL).

TAD Topologically associated domains (TADs) are local chromatin interaction domains, i.e. genomic regions defined by interactome boundaries. TADs represent spatial compartments and are a omnipresent structural feature in the genome [109, 110]. Physical interactions occur more frequently inside a TAD than interactions across TADs. They are cell-type independent and highly conserved.

TAD information can be generated from all-versus-all (also known as Hi-C) chromosome conformation capture experiments. HI-C is an unbiased method to detect chromatin interactions by cross-linking, ligation of spatially adjacent chromatin segments and subsequent high-throughput sequencing of the purified DNA. Then, the interaction frequency between specific chromosomal loci can be determined by quantifying the ligation products. The term all-versus-all refers to the characteristic of a Hi-C experiment to test all possible pairwise interactions between segments are tested.

CADD The combined annotation dependent depletion (CADD) score is used to assess the relative deleteriousness of small genetic variants in coding and non-coding regions [111]. It combines 63 annotations including conservation metrics,

regulatory information like regions of deoxyribonuclease (DNase) hypersensitivity or transcription factor binding, splicing isoform information like distance to exon-intron boundaries or expression levels and protein-level scores into a variant-by-annotation matrix. After training multiple models on observed and simulated variants by using machine learning techniques, an average of the models was used to derive a "C-score". The C-score was then transformed into a Phred-scaled score ("CADD score") between 1 to 99 by applying the formula $-10 \times log_{10}(variant\_rank/number\_of\_variants)$ to the list of variants sorted by their C-score. Accordingly, a CADD score ≥ 10 indicates that the variant belongs to the 10% most deleterious substitutions in the human genome, a score ≥ 20 indicates that the variant belongs to the 1% most deleterious substitutions.

GWAS Catalog The National Human Genome Research Institute (NHGRI)-European Bioinformatics Institute (EBI) Catalog of published genome-wide association studies (also known as GWAS Catalog) [10] is the most comprehensive resource of published SNP-trait/disease associations [78]. It was founded in 2008 by the NHGRI and since then the number of associations increased to multiple tens of thousands of associations. It is updated on a weekly basis by extracting and curating associations from publications.

[10] https://www.ebi.ac.uk/gwas/

Chapter 2

Aims of the Thesis

There is strong evidence that onset and progression of the inflammatory oral disease periodontitis (PD) is significantly influenced by genetic factors, which are still largely unknown [112, 113]. The main goal of this thesis is the identification of novel human genetic factors predisposing to PD. To accomplish this, I utilized human case-control samples of severe forms of PD, including the worldwide largest sample of the most severe and early-onset form aggressive periodontitis (AgP) and the more moderate and slower progressing but widespread form chronic periodontitis (CP).

The specific aims of this thesis are:

1. preparation and refinement of recently generated genotype and phenotype data of the before mentioned AgP case-control sample
2. identification of novel common genetic risk factors for AgP
3. identification of novel common genetic risk factors for PD using all available AgP and CP samples
4. identification of novel shared genetic risk factors for PD and coronary artery disease (CAD)
5. prioritization of candidate causal genes in the identified loci and to identify or confirm molecular pathways that are involved in PD

The identification of PD genetic factors could significantly contribute to identifying the causes, to improve the understanding of the underlying mechanisms and eventually, to improve diagnosis and therapy.

Chapter 3

Manuscripts

3.1 A genome-wide association study identifies nucleotide variants at SIGLEC5 and DEFA1A3 as risk loci for periodontitis

Authors:	**Munz M**, Willenborg C, Richter GM, Jockel-Schneider Y, Graetz C, Staufenbiel I, Wellmann J, Berger K, Krone B, Hoffmann P, van der Velde N, Uitterlinden AG, de Groot LCPGM, Sawalha A, Direskeneli H, Saruhan-Direskeneli G, Guzeldemir-Akcakanat E, Keceli G, Laudes M, Noack B, Teumer A, Holtfreter B, Kocher T, Eickholz P, Meyle J, Doerfer C, Bruckmann C, Lieb W, Franke A, Schreiber S, Nohutcu RM, Erdmann J, Loos BG, Jepsen S, Dommisch H, Schaefer A
Year:	2017
Journal:	*Human Molecular Genetics*
PMID:	28449029
doi:	10.1093/hmg/ddx151
Supplements:	https://academic.oup.com/hmg/article/26/13/2577/3755421

3.1.1 Outline

I conducted a GWAS using German (Ger) and Dutch (NL) case-control samples of aggressive periodontitis (AgP) comprising 851 cases and 6,580 controls after quality control (**Figure 3.1**). Sixteen lead SNPs of the discovery stage passed the pre-defined selection criteria ($P_{AgP} < 10^{-5}$, $P_{AgP} < P_{AgP-Ger}$, $P_{AgP} < P_{AgP-NL}$, $P_{AgP} < 10^{-3}$ for proxy SNPs [$r^2 > 0.8$]) and were validated in a German sample of severe forms of chronic periodontitis (CP) comprising 993 cases and 1,419 controls. Of the sixteen SNPs, two SNPs and their proxy SNPs were not found in the

validation sample. Of the remaining 14 SNPs, two SNPs at *SIGLEC5* (sialic acid binding Ig-like lectin 5) on chromosome (chr) 19q13.41 and a chromosomal region downstream of the *DEFA1A3* locus (defensin alpha 1-3) on chr8p23.1, showed nominal significant association ($P < 0.05$) with CP. At *DEFA1A3*, an additional SNP in low linkage disequilibrium (LD, $r^2 = 0.28$) with the lead SNP, showed nominal association with CP. Subsequently, the SNPs were replicated in a Turkish sample of AgP comprising of 223 cases and 564 controls. In the replication stage, the three SNPs at *SIGLEC5* and *DEFA1A3* showed consistent effect directions with the previous stages, however none of them could be successfully replicated and only the lead SNP at *SLC1A3* reached nominal significance. In the pooled analysis combining discovery, validation, replication samples, three SNPs reached genome-wide significance: rs4284742-G at *SIGLEC5* ($P = 1.09 \times 10^{-8}$, $OR = 1.34$, $95\%CI = 1.21$–1.48); rs2738058-T at *DEFA1A3* ($P = 5.48 \times 10^{-10}$, $OR = 1.28$, $95\%CI = 1.18$–1.38); and rs2978951-A at *DEFA1A3* ($P = 2.06 \times 10^{8}$, $OR = 1.25$, $95\%CI = 1.16$–1.35).

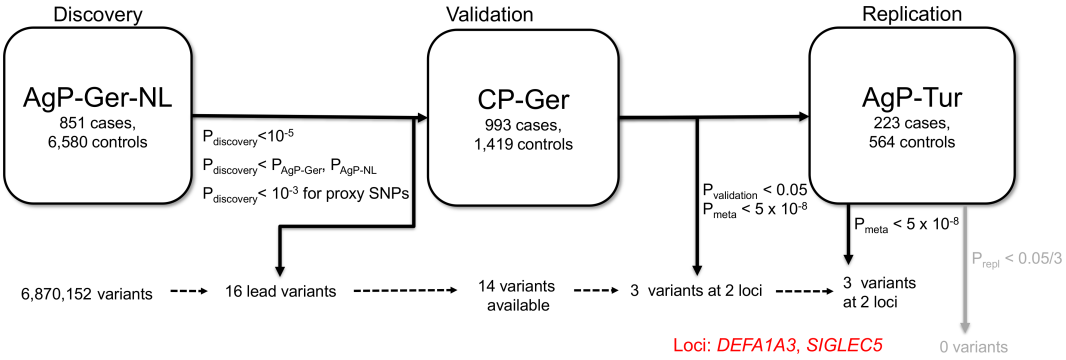

Figure 3.1: Aggressive periodontitis GWAS analysis pipeline. The GWAS analysis pipeline is subdivided into discovery, validation and replication steps. (Grey text) Of the three SNPs that were taken into replication, none could be replicated. However sample pooling resulted into genome-wide significant associations for all three SNPs in two loci (red text). AgP: aggressive periodontitis; CP: chronic periodontitis; Ger: German; NL: Dutch; Tur: Turkish;

Further analysis of the 17 distinct haplotype blocks (16 plus one additional at *DEFA1A3*) suggested an association with AgP for expression quantitative trait loci (eQTLs) indicating regulatory effects for various SNPs. The lead SNP at *SIGLEC5* showed an effect on the transcription level of *SIGLEC5* with $P = 7.7 \times 10^{-14}$ in blood, indicating *SIGLEC5* to be the affected gene of this association. *SIGLEC5* is expressed in various myeloid immune cells and classified as an inhibitory receptor with the potential to mediate tyrosine phosphatases SHP-1/-2 dependent signaling. It is a member of the human CD33-related siglecs and is broadly expressed in various myeloid cells of the innate immune system and in B cells. Experimental data

showed, that *SIGLEC5* efficiently inhibits calcium fluxing, which is mediated by *FCER1G*, one of the sixteen loci that passed are pre-assigned selection criteria in the discovery stage. The lead SNP at *FCER1G* also shows an effect on the expression of *FCER1G* with $P = 9.16 \times 10^{-16}$ in blood. At *DEFA1A3*, two proxy SNPs of the lead SNP show eQTL effects on *DEFA4* with $P = 3.9 \times 10^{-5}$ and $P = 1.7 \times 10^{-3}$ in blood. Alpha defensins are antimicrobial peptides with expression in neutrophils and mucosal surfaces and a role in phagocyte-mediated host defense.

Taken together, this study identifies the first shared genetic risk loci of AgP and CP with genome-wide significance and highlights the role of innate and adaptive immunity in the etiology of periodontitis. Further, the genes *FCER1G* and *SLC1A3* are suggested as promising candidates for future association studies in larger case-control samples.

3.1.2 Contribution

In my role as bioinformatician, I took over the *in silico* work including the creation of an analysis plan and the statistical analyses. The work started with viewing the available datasets and defining an appropriate analysis strategy. Then, I performed the pre-imputation quality control (QC), the genotype imputation and the post-imputation QC of the aggressive periodontitis samples. Subsequently, I collected data of the other case-control samples and ran harmonization procedures in order to analyze all the data together. Finally, I applied statistical measures and tests by using state-of-the-art GWAS methods to generate and interpret the results.

All this work also involved the implementation of a QC and analysis pipeline and the programming of multiple scripts for data integration purposes, calculation of statistics and visualizations. The data integration step also included the fetching and preparation of annotations from the public domain. In addition, data management tasks including storage and handling of large datasets as well as the execution of the pipeline on a computing cluster which was also done by myself.

After the results were generated, I wrote the first version of the manuscript and submitted it to the respective journal. Subsequently, I answered the reviewer comments in a rebuttal. For both the manuscript and rebuttal I got support by my supervisor Arne Schäfer regarding specific questions on the phenotype of periodontitis.

Other authors that are listed on this manuscript were principal investigators or contributed to the work by establishing the included case-control samples, reviewing the manuscript or giving significant input in discussions.

3.2 Meta-analysis of genome-wide association studies of aggressive and chronic periodontitis identifies two novel risk loci

Authors: **Munz M**, Richter GM, Loos BG, Jepsen S, Divaris K, Offenbacher S, Teumer A, Holtfreter B, Kocher T, Bruckmann C, Jockel-Schneider Y, Graetz C, Ahmad I, Staufenbiel I, van der Velde N, Uitterlinden AG, de Groot LCPGM, Wellmann J, Berger Klaus, Krone B, Hoffmann P, Laudes M, Lieb W, Franke A, Erdmann J, Dommisch H, Schaefer AS

Year: 2018

Journal: *European Journal of Human Genetics*

PMID: 30218097

doi: 10.1038/s41431-018-0265-5

Supplements: https://www.nature.com/articles/s41431-018-0265-5

3.2.1 Outline

To obtain novel insights into the shared genetic etiology and the underlying molecular mechanisms of the early-onset and more severe form aggressive periodontitis (AgP) and the late-onset and less severe from chronic periodontitis (CP), I performed a explorative and a validation meta-analysis by using genome-wide association studies of both phenotypes (**Figure 3.2**). I included two German (Ger) and Dutch (NL) samples of AgP (AgP-Ger: 680 cases and 3,973 controls; AgP-NL: 171 cases and 2,607 controls), a German sample of CP (CP-Ger: 993 cases and 1,419) and two European American (EA) samples having a severe (sev) and moderate (mod) form of CP (958 CP-sev cases, 2,293 CP-mod cases, 1,909 controls).

By default, I applied the fixed effects model. However, for variants with a high degree of heterogeneity, i.e. a P-value of Cochran's Q $P(Q) < 0.05$ and a heterogeneity index $I^2 > 0.5$, I applied the random effects model instead. Genetic variants with $P < 10^{-6}$ in the meta-analysis were considered as suggestively associated with PD, variants with $P < 5 \times 10^{-8}$ were considered as genome-wide significant.

Four variants passed the selection criteria for suggestive association. One of them reached genome-wide significance at intronic of the long intergenic non-coding RNA *LOC107984137* on chromosome (chr) 16, downstream of the gene *SHISA9* (Shisa family member 9; rs729876-T, $P = 9.77^{-9}$, $OR = 1.24$, $95\%CI = [1.15 - 1.34]$). The function of *LOC107984137* is not known yet, however suggestive expression

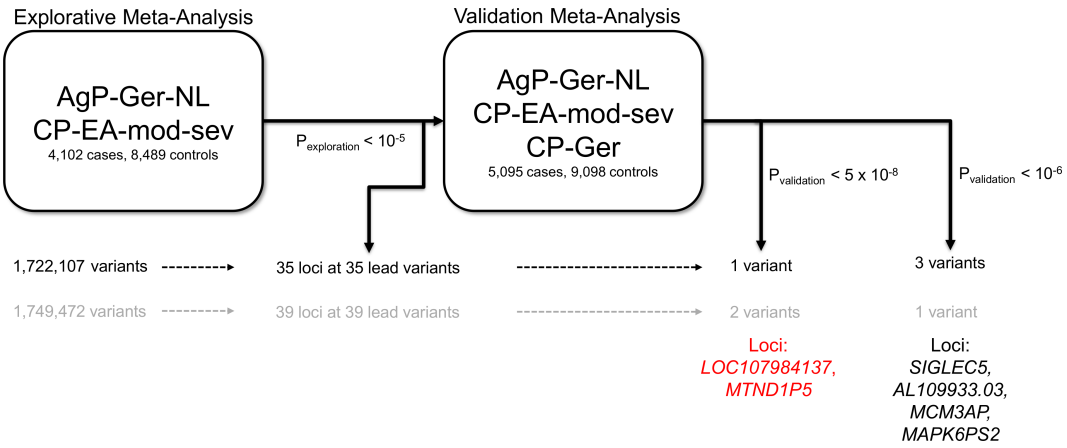

Figure 3.2: Periodontitis meta-analysis pipeline. The pipeline subdivides into an explorative and a validation meta-analysis. (Grey text) The pipeline was re-run again without the smallest sample AgP-NL. (Red text) Two loci reached genome-wide significance. AgP: aggressive periodontitis; CP: chronic periodontitis; Ger: German; NL: Dutch; EA: European-American; mod: moderate; sev: severe;

quantitative loci (eQTL) data indicated tissue specific effects on the expression of the genes *HOXC10* (Homeobox C10) in blood monocytes and *ZC3H7A* (zinc finger CCCH-type containing 7A) and *MYH11* (myosin, heavy chain 11) in the brain. Experimental work suggests that the chimeric protein β/MYH11 inhibits the function of RUNX1 (runt-related transcript factor 1) which plays a role hematopoiesis and bone formation.

Small cohorts such as the Dutch AgP case cohort are generally impaired by random depletion as well as enrichment of risk alleles due to chance effects and can have strong effects on the results, especially if the control sample is large, as is the case with our study. After excluding the Dutch AgP sample, two more variants passed the selection criteria, one variant reached genome-wide significance with $P = 3.69 \times 10^{-9}$ (rs16870060-G, $OR = 1.36$, $95\%CI = [1.23 - 1.51]$) and is located within the pseudogene MTND1P5 on chr8, 13kb downstream of the protein-coding gene *ATP6V1C1* (ATPase H+ transporting V1 subunit C1). eQTL data indicated no *cis*-effect of variants on these genes, but suggest trans-effects on *ARHGEF28* (Rho guanine nucleotide exchange factor 28) in the liver and on *ORM1* (Orosomucoid 1) in monocytes. *ORM1* encodes a key acute phase plasma protein, which is increased due to acute inflammation. Moreover, ORM1 was experimentally shown to interact with PAI-1 (plasminogen activator inhibitor-1) and the binding of PAI-1 to ORM1 results in significant stabilization of its inhibitory activity toward plasminogen activators. *PLG* (Plasminogen) was previously found to be associated with PD and likewise, this locus was among the six associated loci with $P < 10^{-6}$ for the lead SNP.

This study identified novel risk loci of periodontitis, adding to the shared genetic basis of AgP and CP.

3.2.2 Contribution

In my role as bioinformatician, I took over the *in silico* work including the creation of an analyses plan and the statistical analyses. The work started with viewing the available datasets and defining an appropriate analysis strategy. Then, I collected the datasets of the included case-control samples and ran harmonization procedures in order to analyze all the data together in meta-analyses. For samples with overlapping control cohorts I reviewed the literature for methods to account for this issue. Finally, I applied statistical measures and tests by using state-of-the-art GWAS and meta-analysis methods to generate and interpret the results.

All this work also involved the implementation of a QC and analysis pipeline and the programming of multiple scripts for data integration purposes, calculation of statistics and visualizations. The data integration step also included the fetching and preparation of annotations from the public domain. In addition, data management tasks including storage and handling of large datasets as well as the execution of the pipeline on a computing cluster which was also done by myself.

After the results were generated, I wrote the first version of the manuscript and submitted it to the respective journal. Subsequently, I answered the reviewer comments in a rebuttal. For both the manuscript and rebuttal I got support by my supervisor Arne Schäfer regarding specific questions on the phenotype of periodontitis.

Other authors that are listed on this manuscript were principal investigators or contributed to the work by establishing the included case-control samples, reviewing the manuscript or giving significant input in discussions.

3.3 Genome-wide association meta-analysis of coronary artery disease and periodontitis reveals a novel shared risk locus

Authors: **Munz M**, Richter GM, Loos BG, Jepsen S, Divaris K, Offenbacher S, Teumer A, Holtfreter B, Kocher T, Bruckmann C, Jockel-Schneider Y, Graetz C, Munoz L, Bhandari A, Tennstedt S, Staufenbiel I, van der Velde N, Uitterlinden AG, de Groot LCPGM, Wellmann J, Berger K, Krone B, Hoffmann P, Laudes M, Lieb W, Franke A, Dommisch H, Erdmann J, Schaefer AS
Year: 2018
Journal: *Scientific Reports*
PMID: 30209331
doi: 10.1038/s41598-018-31980-8
Supplements: https://www.nature.com/articles/s41598-018-31980-8

3.3.1 Outline

Evidence for a shared genetic basis of association between coronary artery disease (CAD) and periodontitis exists. In this study, we aimed to identify novel genetic risk factors that are shared between CAD and PD, in order to improve the current pathogenic understanding of both diseases and highlight possible common biological underpinnings. I performed a GWAS meta-analysis by using a CAD meta-analysis dataset (60,801 cases vs 123,504 controls) from the "Coronary Artery Disease Genome-wide Replication and Meta-analysis plus The Coronary Artery Disease" (CARDIoGRAMplusC4D) consortium and multiple GWAS datasets for AgP and CP (**Figure 3.3**). More precisely, I included two German (Ger) and Dutch (NL) samples of AgP (AgP-Ger: 680 cases and 3,973 controls; AgP-NL: 171 cases and 2,607 controls), a German sample of CP (CP-Ger: 993 cases and 1,419) and two European American (EA) samples having a severe (sev) and moderate (mod) form of CP (958 CP-sev cases, 2,293 CP-mod cases, 1,909 controls).

By default, I applied the fixed effects model. However, for variants with a high degree of heterogeneity, i.e. a P-value of Cochran's Q $P(Q) < 0.05$ and a heterogeneity index $I^2 > 0.5$, I applied a random effects model instead.

In the discovery stage, I used a German AgP sample and the CARDIoGRAMplusC4D CAD meta-analysis dataset. Apart of the already known shared risk locus at chromosome 9p21.3 (*CDKN2B-AS1* [CDKN2B antisense RNA 1]), two

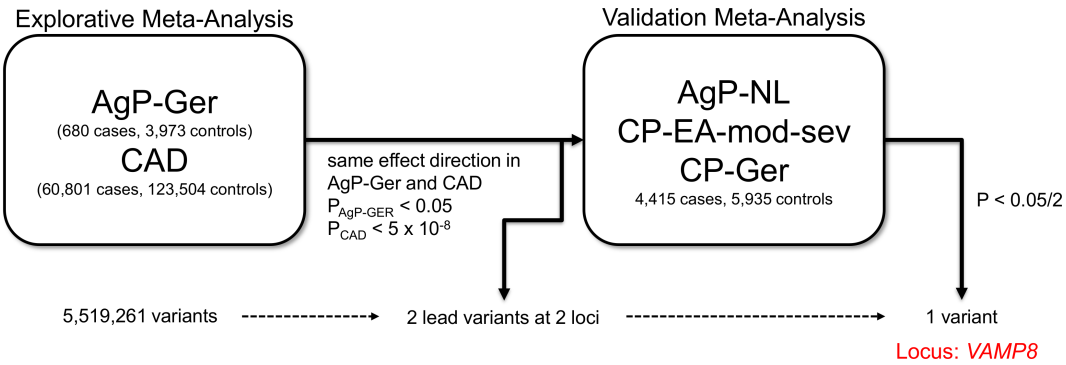

Figure 3.3: Meta-analysis pipeline of periodontitis and coronary artery disease The pipeline subdivides into an explorative and a validation meta-analysis.(Red text) One locus remained significant in the validation step after correcting for multiple testing. AgP: aggressive periodontitis; CP: chronic periodontitis; Ger: German; NL: Dutch; EA: European-American; mod: moderate; sev: severe;

other SNPs at the known CAD risk loci *ADAMTS7* (ADAM metallopeptidase with thrombospondin type 1 motif 7; SNP rs11634042) and *VAMP8* (Vesicle associated membrane protein 8; SNP rs1561198) passed the pre-assigned selection criteria ($P_{AgP-Ger} < 0.05$; $P_{CAD} < 5 \times 10^{-8}$; concordant effect direction) and were replicated in an independent GWAS meta-analysis dataset of PD (4,415 cases vs 5,935 controls). Here, only SNP rs1561198 showed a nominal significant association ($PD_{Replication}$: $P = 0.008$, $OR = 1.09$, $95\%CI = [1.02 - 1.16]$) and could be successfully replicated after adjusting for multiple testing. In the pooled analysis of PD, the association P-value for this SNP refined to $P = 0.0002$ ($OR = 1.11$, $95\%CI = [1.05 - 1.17]$). For the associated haplotype block, allele specific *cis*-effects on *VAMP8* expression were reported. This finding complements results of a previous report addressing the shared molecular mechanisms of both diseases. In this study, a transcriptome-wide shRNA knock-down approach demonstrated that CDKN2B-AS1 and VAMP3 expression is correlated on the RNA and protein level, and a rare variant upstream of VAMP3 (SNP rs17030881) was suggested to be associated with AgP and CAD. This variant as well as another previously reported shared risk loci at PLG (SNP rs4252120) were not re-discovered in this analysis, because they did not pass the pre-defined criterion $P_{CAD} < 5 \times 10^{-8}$. VAMP3 and VAMP8 threshold membrane trafficking in platelets plays an important role in thrombosis and wound healing, processes with established relevance for the etiology of CAD and PD.

Our data adds to the shared genetic basis of CAD and PD and indicates that the observed association of the two disease conditions cannot be solely explained by shared environmental risk factors. We conclude that the molecular pathway shared by CAD and PD involves VAMP8 function, which has a role in membrane vesicular

trafficking, and is manipulated by pathogens to corrupt host immune defense.

3.3.2 Contribution

I took over the *in silico* work including the creation of an analyses plan and the statistical analyses. The work started with viewing the available datasets and defining an appropriate analysis strategy. Then, I collected the datasets of the included case-control samples and ran harmonization procedures in order to analyze all the data together in meta-analyses. For samples with overlapping control cohorts I reviewed the literature for methods to account for this issue. Finally, I applied statistical measures and tests by using state-of-the-art GWAS and meta-analysis methods to generate and interpret the results.

All this work also involved the implementation of a QC and analysis pipeline and the programming of multiple scripts for data integration purposes, calculation of statistics and visualizations. The data integration step also included the fetching and preparation of annotations from the public domain. In addition, data management tasks including storage and handling of large datasets as well as the execution of the pipeline on a computing cluster which was also done by myself.

After the results were generated, I wrote the manuscript and submitted it to the respective journal. Subsequently, I answered the reviewer comments in a rebuttal. For both the manuscript and rebuttal I got support by my supervisor Arne Schäfer regarding specific questions on the phenotype of periodontitis.

Other authors that are listed on this manuscript were principal investigators or contributed to the work by establishing the included case-control samples, reviewing the manuscript or giving significant input in discussions.

3.4 The PF4/PPBP/CXCL5 gene cluster is associated with periodontitis

Authors:	Shusterman A, **Munz M**, Richter G, Jepsen S, Lieb W, Krone B, Hoffman P, Laudes M, Wellmann J, Berger K, Kocher T, Offenbacher S, Divaris K, Franke A, Schreiber S, Dommisch H, Weiss E, Schaefer AS, Houri-Haddad Y, Iraqi FA
Year:	2017
Journal:	*Journal of Dental Research*
PMID:	28467728
doi:	10.1177/0022034517706311
Supplements:	http://journals.sagepub.com/doi/suppl/10.1177/0022034517706311

3.4.1 Outline

Sample sizes are a major limiting factor in the identification of new susceptibility loci in genome-wide association studies. Especially for periodontitis, there is a general scarcity of high-quality phenotype-genotype samples. To overcome this issue, we developed a two stage approach. First, we took advantage of QTL and mRNA data from mice infected with periodontal pathogens to identify promising candidate loci. Then, we ran a candidate gene study on the corresponding human orthologues, by using human GWAS samples of aggressive and chronic periodontitis.

More specifically, the mouse line BALB/cJ which is susceptible for human periodontal pathogens and line A/J which is resistant for the same pathogens, were crossed. Samples of the F_2 generation were then used to run a quantitative trait loci (QTL) data on alveolar bone loss upon infection with periodontal pathogens. Moreover, mRNA sequencing using the Illumina HiSeq platform was applied on four susceptible and four resistant samples of the F_2 generation. The sequencing data was then used to run a differential expression analysis.

Differentially expressed genes (DEGs) resulted from the software tools for differential expression analysis Cuffdiff and DESeq, 251 and 316 DEGs respectively and passed the pre-defined filtering criteria (Fragments per kilobase transcript per million mapped reads [FPKM] > 0 in both conditions, false discovery rate [FDR] < 0.1). DEGs, that occurred in both sets and that were located nearby a QTL, were selected as candidates. The intersection of the DEGs and the QTL data resulted in four loci at the murine genes *Ugt2a1*, *Pf4*, *Sult1d1*, and *Stoml3* which were then mapped to the human orthologous protein-coding genes *UGT2A1*, *PF4*, *STOML3* and the

pseudogene *SULT1D1*.

The loci of the four human orthologous genes +/- 200 kilobases were analyzed for putative variant phenotype associations in three GWAS case-control samples. In the aggressive periodontitis sample of German and Dutch descent, variants at *PF/PPBP/CXCL5* (SNP rs1595009: $P = 1.3 \times 10^{-4}$, odds ratio $[OR]$= 1.32, 95% confidence interval $[CI]$= 1.15 − 1.52) and *UGT2A1-SULT1D1P* (SNP rs146712414, $P = 9.1 \times 10^{-5}$, $OR = 1.34$, $95\%CI = 1.16 - 1.56$) showed an association with $P < 10^{-3}$. These two variants were validated in a European-American sample of moderate chronic periodontitis (1,961 cases, 1,864 controls). Only SNP rs1595009 remained nominally significant ($P = 0.03$, $OR = 1.45$, $95\%CI = 1.01-1.29$). Lastly, we replicated this SNP in a German sample of chronic periodontitis (399 cases, 1,633 controls). In this sample, the SNP couldn't be replicated (i.e. $P > 0.05$) in the additive genetic model. However, when using the recessive model the SNP showed a significant association of $P = 0.03$ ($OR = 1.75$, $95\%CI = 1.06-2.90$). The combined association estimates of all three GWAS samples were $P = 2.9 \times 10^{-5}$ ($OR = 1.2$, $95\%CI = 1.1 - 1.3$). The *PF/PPBP/CXCL5* locus is shared by different disease phenotypes including inflammatory bowel disease.

According to the GTEx (Gene Tissue Expression Project) and the Haploreg database, SNP rs1595009 has eQTL effects on *PF4V1* (platelet factor 4 variant 1; $P = 3.9 \times 10^{-10}$) and *CXCL5* (chemokine [C-X-C motif] ligand 5; $P = 6.5 \times 10^{-6}$) in blood. The identified eQTLs do not exactly correspond with the murine DEG *Pf4*, although the chemokine gene *PF4V1* gene is paralogous to *PF4* and shares its antiangiogenic and anticoagulant effects. In general, eQTLs are tissue-specific and one reason could be the differing tissues. The reported human eQTL effects were derived from human blood, whereas the DEGs were derived from murine gingival tissues. eQTL data from human gingiva do not currently exist. In addition, eQTL effects could differ between humans and mice.

In this work, we identified a new susceptibility locus for periodontitis at *PF/PPBP/CXCL5*. Moreover, by incorporating a mouse model in our search for new susceptibility loci in humans, we provide an efficient approach to identify and prioritize loci that are likely to be missed by traditional GWAS.

3.4.2 Contribution

After the work with the mouse model (mRNA sequencing, QTL analysis, differential expression analysis) was finished by collaborators, I received the filtered set

of differentially expressed mouse genes. I mapped them to the corresponding orthologous gene in human. After determinating the genetic regions in the human genome, I extracted the relevant SNP data from the summary statistics of the three utilized GWAS samples and ran the candidate gene study (step 1: hypothesis generation; step 2: validation; step 3: replication) and the eQTL analysis. Moreover, I contributed to the conception, study design, drafted and critically revised the manuscript.

3.5 GWAS for Interleukin-1β levels in gingival crevicular fluid identifies IL37 variants in periodontal inflammation

Authors:	Offenbacher S, Jiao Y, Kim SJ, Marchesan J, Moss KL, Jing L, Divaris K, Bencharit S, Agler CS, Morelli T, Zhang S, Sun L, Seaman WT, Cowley D, Barros SP, Beck JD, **Munz M**, Schaefer AS, North KE
Year:	2018
Journal:	*Nature Communications*
PMID:	30206230
doi:	10.1038/s41467-018-05940-9
Supplements:	https://www.nature.com/articles/s41467-018-05940-9

3.5.1 Outline

Cytokine levels in the gingival crevicular fluid (GCF) reflects the microbial activation of the host's immune response in the periodontal tissue. IL1B is a member of the interleukin-1 (IL1) cytokine family and is a known GCF biomarker for periodontal disease progression. Twin-studies have shown, that an estimated 86% of the variance of the IL1B levels can be explained by the underlying genetic architecture, meaning that the IL1B secretion is highly heritable.

To find out how GCF IL1B levels are genetically controlled, a GWAS was performed for this continuous phenotype using 4,910 European-Americans of the dental "The Atherosclerosis Risk in Communities" (ARIC) dataset. Linear regression with an additive genetic model and the covariates age, sex, examination center, and ancestry were used to discover putative associations for 656,292 high-quality genotyped SNPs. 72 SNPs passed the genome-wide significant level ($P < 5 \times 10^{-8}$) and revealed a quantitative trait locus (QTL) at *IL37* within the *IL1* gene cluster. IL37 plays a role in the inhibition of the innate immune response and inflammatory reactions.

The *IL1* gene cluster was closely examined by contrasting quartiles of the GCF IL1B level distribution by using logistic regression. In this analysis, it was shown that elevated IL1B levels (3rd vs. 1rd and 2nd quartiles) are mainly associated with variants at *IL1B*, whereas high IL1B levels (4th vs. 1rd, 2nd and 3rd quartiles) are mainly associated with variants at *IL37*. At *IL37*, two haplotype blocks in low LD ($r^2 = 0.23$) showed association, tagged by the missense variants rs3811046 ($P = 3.3 \times 10^{-22}$) and rs2708943 ($P = 4.2 \times 10^{-7}$). Prediction of mRNA structure

of *IL37* using the "Predict a Secondary Structure Web Server" indicated that both wildtype and alternative allele of rs3811046 result in similar secondary structure. On the other hand, alternative alleles for SNP rs2708943 or both SNPs rs3811046 and rs2708943 together, resulted in a structure which significantly differs from the wildtype.

The associations of SNP rs3811046 with high levels of IL1B were replicated in an independent cohort of 143 subjects. Genotypes and GCF levels were measured by using pyrosequencing and multiplex immunoassay, respectively. The comparison of the wildtype genotype and the genotype consisting of two alternative alleles for SNP rs3811046 showed a statistically significant elevation level of IL-1β for the latter genotype which is concordant with the GWAS results. This could also be shown for dendritic cells. An analogous comparison of SNP rs3811046 regarding the elevation level of IL8, one of six other mediators of the innate immune response, that were also tested in the replication step, was also statistically significant. Moreover, SNP rs3811046 showed associations with aggressive periodontitis ($OR = 1.12$, $95\%CI = [1.01-1.26]$), severe chronic periodontitis ($OR = 1.5$, $95\%CI = [1.12-2.00]$) and 10-year incident tooth loss (≥ 3 teeth, relative risk [RR]$= 1.33$, $95\%CI = [1.09-1.62]$).

In follow-up analyses, *IL37* and the two identified genetic variants were functionally characterized in several assays. Relative expression of five *IL37* splicing isoforms were assessed in inflamed human gingival tissue from PD patients by quantifying mRNA levels using real time polymerase chain reaction (qPCR). *IL37b*, the isoform which leads to the longest amino acid (aa) sequence (218aa) was found to be predominantly expressed. Moreover, isoform *IL37b* was significantly higher expressed in PD subjects than in a control sample.

In vivo effects of recombinant human wildtype IL37b (WT rhIL37b) were investigated in a murine periodontitis model. Comparison of IL1B level in gingival tissue in a group of intraperitoneal injected mature WT rhIL37b and a control group by qPCR and immunohistochemistry showed a significant decrease of mRNA and protein level. On the basis of this, rhIL37b might suppress IL1B production in inflamed gingival tissue.

Protein and mRNA levels were measured for WT IL37b and the two missense SNPs in human embryonic kidney (HEK293T) cells that were transfected with corresponding plasmids using qPCR and Western blotting. Protein levels were lower for the single and combined SNPs than for WT. mRNA levels were much higher for WT and rs3811046 than for rs2708943 and both SNPs combined suggesting an impairment

of the mRNA by rs2708943 or a combination of rs3811046 and rs2708943 which results into reduced protein synthesis. HEK293T were co-transfected with caspase-1 to investigate the effects of rs3811046 on the cleavage of the pro form of IL37b. The results of this analysis suggested that rs3811046 leads to aberrant production and secretion of IL37.

The work reveals a genome-wide significant association between SNP rs3811046 at *IL37* IL1B levels in the gingival crevicular fluid. SNP rs3811046 leads to an impaired production and cleavage of IL37 which in turn increases the expression of IL1B. Further, this variant can be significantly associated with aggressive and chronic periodontitis. A second variant at *IL37*, SNP rs2708943, dramatically reduces mRNA expression of *IL37*.

3.5.2 Contribution

I validated the association of SNP rs3811046 at *IL37* with the phenotype "gingival crevicular fluid IL-1β levels" that was discovered in the first part of the manuscript, in the German and Dutch case-control samples of aggressive periodontitis. This involved the extraction of the corresponding genotypes from the overall genotype files, the testing for association and the combination of the results in the German and Dutch samples in a subsequent meta-analysis. Finally, I summarized my results as well as materials and methods I used to be included in the manuscript.

Chapter 4

Discussion

In my work, I aimed to elucidate the human genetic risk factors of periodontitis (PD) to enhance the current etiological concept and to provide new therapeutic targets for the prevention and treatment of PD. I used three approaches to identify novel risk loci for PD. In my first approach (chapter 3.1 on page 19), I exploited the unique characteristics of the rare and very severe phenotype of aggressive periodontitis (AgP), which is supposed to be determined by genetic factors to a stronger degree compared to less severe forms of PD. To this end I used a German and a Dutch AgP sample (851 cases, 6,580 controls) in the discovery stage. The resulting variants were then validated in a German chronic periodontitis (CP) sample (993 cases, 1,419 controls) and replicated in a Turkish AgP sample (223 cases, 564 controls). The CP sample was selected for validation to complement the small Turkish sample and the lack of other AgP samples. I was able to identify novel genetic risk loci at *SIGLEC5* (Sialic acid binding Ig-like lectin 5) on chromosome (chr) 19q13.41 and at *DEFA1A3* (Defensin alpha 1 and alpha 3) on chr8p23.1. Both discovery loci could not be confirmed in the individual CP and Turkish AgP samples, but reached but reached genome-wide significance after pooling them together, providing solid statistical evidence for the relevance of these loci in the etiology of PD.
In the second approach (chapter 3.2 on page 35), I pooled all available genome-wide association scans of severe forms of periodontitis, i.e. samples of AgP and CP, in a meta-analysis to maximize the statistical power. This comprised altogether of 5,095 cases and 9,908 controls, and resulted in two genome-wide associated loci at *MTND1P5* (Mitochondrially encoded NADH:ubiquinone oxidoreductase core subunit 1 pseudogene 5) on chr8q22.3 and at the long intergenic non-coding RNA *LOC107984137* on chr16.
For my third approach (chapter 3.3 on page 51), I sought to explore the shared genetic basis of PD and coronary artery disease (CAD) by identifying known CAD

loci that also show an association with PD. In the discovery step, I compared a large CAD meta-analysis sample comprising of 60,801 cases and 123,504 controls and the aforementioned German AgP sample for overlapping loci. Subsequently, the variants with the best shared association were validated in the aforementioned Dutch AgP sample and multiple CP samples (4,415 cases and 5,935 controls). An association at *VAMP8* (Vesicle associated membrane protein 8) on chr2p11.2 could be successfully replicated after correcting for multiple testing.
Moreover, I contributed to two other projects that aimed to identify new risk factors for PD. The first project (chapter 3.4 on page 65) consisted of a RNA sequencing based differential expression analysis in mice regarding the susceptibility to infection with oral bacterial pathogens, and a subsequent candidate gene study of the differentially expressed orthologous genes in human using GWAS data of the German AgP sample and a European-American (EA) CP sample (1,961 cases and 1,864 controls). This project identified an association of the gene cluster *PF4* (platelet factor 4)/*PPBP* (pro-platelet basic protein)/*CXCL5* (C-X-C motif chemokine ligand 5) on chr4q13.3 with PD.
The second project (chapter 3.5 on page 77) consisted of a GWAS regarding IL1B (Interleukin 1 beta) expression in gingival crevicular fluid in an EA sample ($n = 4910$) and a validation with CP samples and the German and Dutch AgP samples. This project revealed a novel PD association at *IL37* (Interleukin 37) on chr2q14.1.

My three approaches identified multiple new risk loci and suggest causal haplotype blocks which are proxied by their corresponding top variants. By virtue of the linkage disequilibrium structure in the human genome, a haplotype block is generally defined by tens to hundreds of closely correlated variants. Accordingly, the top variants of the novel PD risk loci are also tighlty linked to multiple other variants. Therefore, further analyses should be carried out to identify the putative causative variant(s) underlying the associations at these loci. Commonly used methods include *in silico* prediction tools, fine-mapping of disease-associated regions followed by experimental validation [114]. To unravel a true causative variant from other putative causal variants in a fine-mapping approach, very large sample sizes are needed. Moreover, experimental validation can include complex experimental designs. However, a more relevant biological question is to assign the causal gene(s) to every GWAS locus. Similarly to the causative variants, a risk locus can also harbor multiple causal genes. Pinpointing these and translating the identified genetic variants into disease mechanisms and are essential for improving diagnosis and therapy. For this reason, I have focused my post-GWAS analyses on the prioritization of candidate genes.

Similarly to the identification of a risk variant, the identification of a causal gene is not straight forward. The closest gene to the lead variant, i.e. the variant with the lowest P-value, is not always the causal gene [115]. This can be explained by the chromatin looping, through which DNA sequences such as regulatory enhancer elements harboring a causal variant and gene promotors, can be spacially placed very close to each other, thus enabling interactions of more distant DNA elements [116]. Hence, to prioritize candidate causal genes, I combined genome-wide knowledge by integrating multiple datasets of the public domain including scientific publication as well as information on topological associated domains and expression quantitative trait loci.

My integrated dataset findings suggested multiple candidate causal genes in *cis* and *trans* of the corresponding risk harboring haplotype blocks. These genes are involved in innate and adaptive immunity (*SIGLEC5*, *DEFA1A3*, *ORM1* (Orosomucoid 1) on chr 9q32, *VAMP8*), wound healing (*ORM1*, *VAMP8*), hematopoiesis (*MYH11* (Myosin, heavy chain 11, smooth muscle) on chr16p13.11, *RUNX1* (Runt-related transcription factor 1) on chr21q22.12, *VAMP8*) and bone homeostasis (*SIGLEC5*, *MYH11*, *RUNX1*). The importance of these molecular mechanisms in the etiology of PD is also supported by putative but not yet validated risk loci, that suggested an association with our analyses (*FCER1G* [Fc fragment of IgE, high affinity I, receptor for; gamma polypeptide] on chr 1q23.3, *SLC1A3* [Solute carrier family 1, member 3] on chr5p13.2) and is reflected in the pathogenesis of AgP and CP which is characterized by severe inflammation, eventually leading to extensive resorption of the alveolar bone. Especially in the case of AgP, the inflammation is often not accompanied by obvious signs of a particular pathogenic burden such as increased oral plaque or reddening and swelling of the gums, which could reflect an aberrant reactivity of the immune system to yet unknown triggers or intolerance to resident oral bacteria.

It should be noted that the PD phenotypes of the sampels that were used in this thesis, are based on different definitions and classifications. For example, the AgP samples consisted of cases with inclusion critera ≥ 2 affected teeth with $\geq 30\%$ bone loss and a population representative control cohort; or in the German CP sample where cases and controls were defined by contrasting subjects within the first versus the third tertile of proportion of proximal sites with attachment loss (AL) $\geq 4mm$. However, heterogeneity of PD phenotype definition is not unusual because there doesn't exist single general accepted classification for PD, yet. This indicates that no existing definition gives a complete description of the full spectrum of phenotypes. We consider that the various PD manifestations are part of a large range of similar

conditions that are attributed to the effects of different combinations of genetic variants and environmental factors. In this view, AgP which is the most severe and early-onset form of PD, would be affected to a higher degree by genetic risk factors, whereas CP, which is characterized by a late age of disease-onset, would be affected to a lower degree by genetic risk factors, but would be affected stronger by the accumulating negative effects of environmental factors and ageing. Yet, both forms essentially differ in the speed of progression and we were aware, that increased phenotype-heterogeneity could diminish the gain of statistical power from larger sample sizes by diluting the average genetic effects [117]. However, since AgP and CP do not differ in the clinical picture of alveolar bone loss, we rated the gain of statistical power to identify shared genetic risk factors as being superior. We think, that PD is best defined by percentage of alveolar bone loss in response to periodontal inflammatory processes. Moreover, during the last decade, the available samples were constructed independently. Consequently, we combined samples of both PD forms and different phenotype definitions in my analyses.

Apart of the novel risk loci, my analyses also provided the opportunity to confirm the previously identified GWAS risk loci for PD *VAMP3*, *ANRIL* (CDKN2B antisense RNA 1) on chr9p21.3, *GLT6D1* (Glycosyltransferase 6 domain containing 1) on chr9q34.3, *PLG* (Plasminogen-like B1) and *PF4/PPBP/CXCL5*. However, it must be noted that the samples that were used for detecting these loci are not independent from the samples I used in my work. Indeed, there is an overlap of up to 32% of the cases and 54% of the controls in the pooled PD sample. Regarding the AgP samples, either subsets (approximatelly 50% of the cases) or the complete samples were used in these previous studies. The confirmation of the loci *VAMP3* was not possible, because the corresponding association signal belonged to a rare haplotype block with a minor allele frequency (MAF) $< 5\%$. However, all my analyses were restricted to common variants with a MAF $> 5\%$ due to statistical power considerations. Variants that did not pass this threshold were filtered out before testing for association. Moreover, a common variant at *VAMP3* which was previously identified using subjects of the European-American CP sample, a sample which was included in my second and third approach, couldn't be identified either. A possible reason could be that the common variant was previously found to be associated with periodontal microbiota instead of PD [54]. In the discovery of my first approach, all the remaining associations reached at least nominal significance ($P \leq 0.5$) in German AgP, but none of the loci were among the suggestive association at the significance level $P < 5 \times 10^{-5}$. The preceding identification of the association at *ANRIL*, which has the largest effect size among all known PD associations, was

carried out using a recessive genetic model. However, in all my analyses I applied the additive model which, on the one side, is more generic, but on the other side, as a matter of course, is less powerful to detect risk factors with a recessive nature than a recessive model. In addition, although the *ANRIL* locus reached nominal significance in the German AgP sample, the overall effect in AgP is strongly diminished by the Dutch AgP sample, because the effect allele is enriched in the cases only for the heterozygous genotype and thus compensates for the overall effect. The loci *GLT6D1* and *PLG* did also not show up in the first approach, either because of stochastic variability of the allele distribution or due to lack of statistical power. In the second approach, the *PLG* locus occurred in the top list of associated variant and confirms this locus as a shared risk factor of AgP and CP. *ANRIL* and *GLT6D1* were not in the list of suggestive loci, underlining that these loci harbor AgP risk only and a lack of statistical power in my first approach. In my third approach *ANRIL* re-appeared as top hit in the discovery. The locus at *PLG*, which was the second known shared risk locus of CAD and PD, didn't come up which could be attributed to the lack of statistical power as well. The *PF4/PPBP/CXCL5* gene cluster was also not present in the list of suggestive association signals in all three approaches. It was previously identified in a candidate gene study, which included an additional CP sample in the replication stage that was only genotyped for the specific region of the *PF4/PPBP/CXCL5* cluster. Thus, the non-presence might be due to statistical power issues. Still, this gene cluster re-appeared as a risk locus in the candidate gene study, which was performed in one of the two projects I contributed to.

The new findings in this work underline the shared genetic basis between AgP and CP. This may point to a high genetic correlation between AgP and CP and indicate that they are indeed sub-phenotypes belonging to the spectrum of a single disease. SNP-based methods exist to measure the extent of the shared genetic basis across the whole genome by estimating the genetic correlation [118, 119]. However, they require large sample sizes to deliver meaningful results. For example the tool LD Hub [1] recommends to use sample sizes with > 5000 patients, which exceeds the size of the AgP size by more than fivefold. In addition, genome-wide summary statistics were not available to us for all samples (German CP samples, Turkish AgP). Thus, serious future study attempts should be postponed, when larger sample sizes are available. The same applies for heritability estimations and the confirmation of the assumption that AgP is more heritable than CP. Yet, it should be noted that a greater heritability is observed for many early-onset phenotypes compared to late onset phenotypes in numerous diseases and traits, e.g. diabetes, myocardial

[1] http://ldsc.broadinstitute.org/ldhub/

infarction and obesity.

The results of the presented analyses more than doubled the number of validated PD risk loci from five to eleven loci which could be considered as low in comparison to other complex inflammatory diseases. For example, 143 risk loci are already known for type 2 diabetes [34] and even more ($n = 163$) are known for CAD (reviewed in [43]). But even for these diseases the proportion of heritability that can be explained by the validated risk loci is small. This issue is also referred to as the "missing heritability" problem which could be attributed, inter alia, to the high false-negative rate due to lack of statistical power and to genetic interations, which creates a "phantom heritability" [120]. To identify further risk variants, new studies with even larger sample sizes are required. Such larger samples sizes are needed because the effect of yet unknown risk variants decreases with the number of risk variants identified, requiring more power to detect them. To-date, the locus with the lowest effect size for CAD is at *CETP* (Cholesteryl ester transfer protein, plasma) on chr16q13 showing an Odds Ratio (OR) of 1.03 for SNP rs1800775 [121]. This study comprises more than 70,000 cases and 120,000 controls. In comparison, the risk variant of PD with the lowest effect size is at *VAMP8* with *OR* of 1.11 for SNP rs1561198. Interestingly, the lower the OR gets for which studies are carried out with adequate sample sizes, the more new risk variants are identified, until a plateau phase is reached [66].

If this also applies for other diseases, it means, that a lot more loci contribute to the risk of a complex disease than previously expected. In line with this, Boyle et al. published a paper that prompted discussion, suggesting that nearly all genes contribute to a disease condition, i.e. also genes that are not of direct relevance to a disease, refering to it as the "omnigenic" model [122]. According to the model, most of the genes are contributing indirectly to a disease and show relatively small effect sizes. These genes are classified as being "peripheral" and are thought to show a large amount of pleitropy. By contrast, genes, that play a more direct role in a disease harbor rare disease-specific variants with relatively large effects sizes. Hence, those genes are classified "core" genes. The inolvement of a large number of genes in a complex disease seems to be generally accepted by the scientific community. Large numbers of variants have already been adopted in polygenic risk score calculations (PRS) and proven to be superior to the classical PRS comprising only known and validated risk variants [123]. However, the concept of only a few "core" genes contributing to a disease is critically discussed [124], because large exome or genome sequencing studies yet failed to confirm this for common diseases [125–127]. Wray et al. hypothesize that differing phenotypes between common variant and rare

variant associated traits or diseases may be a possible explanation why these studies have failed confirmation. In other words, to identify rare variants with large effect sizes, a division into sub-phenotypes would be required, e.g. by the stratification of samples into lethality, severity, age-of-onset, etc. On the other hand, common diseases are actually uncommon in a population, i.e. most people are healthy. This indicates an inherent robustness in the biological system which is why an etiology of many core genes cannot be excluded. This means an indistinguishability between peripheral and core genes. A disease that impacts only a small fraction of the population with a genetic architecture of many risk loci with similar effect sizes could then be explained by a high non-linear relationship between probability of a disease and burden of risk alleles, implying that polygenic disease is non-additive on the disease scale but rather caused by interacting genetic effects.

In conclusion, with the GWAS approches presented in this thesis, I identified and validated five novel susceptibility loci for PD by combined analyses of multiple case-control samples of AgP and CP and a large CAD sample. Four of these are associated at genome-wide significance level, providing statistical evidence for the relevance of these loci in the disease etiology. The fifth locus represents a novel shared risk locus between PD and CAD. Moreover, I contributed to a project in which an additional risk locus was found for a PD-related phenotype and validated for both AgP and PD. The suggested candidate causal genes highlight mechanisms in inflammation and bone metabolism to be important features in the etiology of PD. To validate the candidate genes and to identify the causative variants, further functional analyses are required. Furthermore, we could replicate some of the previously reported loci. It is likely that statistical power limitations kept us from replicating all relevant signals and finding more true positive signals. Larger samples are needed to fully validate all the signals that were suggested by our analyses and to identify further risk variants. As a supplement to future works, studies on rare, extreme phenotypes in families could be conducted. In my thesis, the analyses of rare variants as well as the X chromosome were not included. With the new X chromosome imputation feature available on the *Sanger Imputation Server*[2] and increasing sample sizes, a consideration of these dimensions might be worthwile in further investigations to enhance the understanding of PD. Until then, other approaches such as compound heterozygosity [128] analyses could be promising alternative enabling the increase of power at constant sample sizes.

[2]https://imputation.sanger.ac.uk

Bibliography

[1] C. A. J. Janeway et al. *Immunobiology 5 : the immune system in health and disease*. Garland Science, 2001, p. 732.

[2] M. F. Flajnik and M. Kasahara. "Origin and evolution of the adaptive immune system: genetic events and selective pressures". In: *Nature Reviews Genetics* 11.1 (2010), pp. 47–59. DOI: 10.1038/nrg2703. PMID: 19997068.

[3] M. Rücklin et al. "Development of teeth and jaws in the earliest jawed vertebrates". In: *Nature* 491.7426 (2012), pp. 748–751. DOI: 10.1038/nature11555.

[4] P. M. Bartold and T. E. Van Dyke. "Periodontitis: a host-mediated disruption of microbial homeostasis. Unlearning learned concepts." In: *Periodontology 2000* 62.1 (2013), pp. 203–17. DOI: 10.1111/j.1600-0757.2012.00450.x. PMID: 23574467.

[5] B. S. Michalowicz et al. "Evidence of a Substantial Genetic Basis for Risk of Adult Periodontitis". In: *Journal of Periodontology* 71.11 (2000), pp. 1699–1707. DOI: 10.1902/jop.2000.71.11.1699. PMID: 11128917.

[6] W. Marcenes et al. "Global burden of oral conditions in 1990-2010: a systematic analysis." In: *Journal of dental research* 92.7 (2013), pp. 592–7. DOI: 10.1177/0022034513490168. PMID: 23720570.

[7] P. I. Eke et al. "Prevalence of Periodontitis in Adults in the United States: 2009 and 2010". In: *Journal of Dental Research* 91.10 (2012), pp. 914–920. DOI: 10.1177/0022034512457373. PMID: 22935673.

[8] R. Ide, T. Hoshuyama, and K. Takahashi. "The Effect of Periodontal Disease on Medical and Dental Costs in a Middle-Aged Japanese Population: A Longitudinal Worksite Study". In: *Journal of Periodontology* 78.11 (2007), pp. 2120–2126. DOI: 10.1902/jop.2007.070193. PMID: 17970678.

[9] W. Nesse et al. "Periodontal inflamed surface area: quantifying inflammatory burden". In: *Journal of Clinical Periodontology* 35.8 (2008), pp. 668–673. DOI: 10.1111/j.1600-051X.2008.01249.x. PMID: 18564145.

[10] C. Susin, A. N. Haas, and J. M. Albandar. "Epidemiology and demographics of aggressive periodontitis". In: *Periodontology 2000* 65.1 (2014), pp. 27–45. DOI: 10.1111/prd.12019. PMID: 24738585.

[11] N. Kassebaum et al. "Global Burden of Severe Periodontitis in 1990-2010". In: *Journal of Dental Research* 93.11 (2014), pp. 1045–1053. DOI: 10.1177/0022034514552491.

[12] B. Holtfreter et al. “Standards for reporting chronic periodontitis prevalence and severity in epidemiologic studies”. In: *Journal of Clinical Periodontology* 42.5 (2015), pp. 407–412. DOI: 10.1111/jcpe.12392.

[13] W. M. M. Jenkins and P. N. Papapanou. “Epidemiology of periodontal disease in children and adolescents”. In: *Periodontology 2000* 26.1 (2001), pp. 16–32. DOI: 10.1034/j.1600-0757.2001.2260102.x.

[14] R. T. Demmer and P. N. Papapanou. “Epidemiologic patterns of chronic and aggressive periodontitis”. In: *Periodontology 2000* 53.1 (2010), pp. 28–44. DOI: 10.1111/j.1600-0757.2009.00326.x.

[15] D. F. Kinane, H. Shiba, and T. C. Hart. “The genetic basis of periodontitis”. In: *Periodontology 2000* 39.1 (2005), pp. 91–117. DOI: 10.1111/j.1600-0757.2005.00118.x.

[16] K. Divaris et al. “Exploring the genetic basis of chronic periodontitis: a genome-wide association study.” In: *Human molecular genetics* 22.11 (2013), pp. 2312–24. DOI: 10.1093/hmg/ddt065. PMID: 23459936.

[17] S. Offenbacher et al. “Genome-wide association study of biologically informed periodontal complex traits offers novel insights into the genetic basis of periodontal disease.” In: *Human molecular genetics* 25.10 (2016), pp. 2113–2129. DOI: 10.1093/hmg/ddw069. PMID: 26962152.

[18] D. F. Kinane, P. G. Stathopoulou, and P. N. Papapanou. “Periodontal diseases”. In: *Nature Reviews Disease Primers* 3 (2017), p. 17038. DOI: 10.1038/nrdp.2017.38. PMID: 28805207.

[19] P. N. Papapanou et al. “Periodontitis: Consensus report of workgroup 2 of the 2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions”. In: *Journal of Periodontology* 89 (2018), S173–S182. DOI: 10.1002/JPER.17-0721. PMID: 29926951.

[20] J. G. Caton et al. “A new classification scheme for periodontal and peri-implant diseases and conditions - Introduction and key changes from the 1999 classification”. In: *Journal of Clinical Periodontology* 45 (2018), S1–S8. DOI: 10.1111/jcpe.12935. PMID: 29926489.

[21] J. Kim and S. Amar. “Periodontal disease and systemic conditions: a bidirectional relationship.” In: *Odontology* 94.1 (2006), pp. 10–21. DOI: 10.1007/s10266-006-0060-6. PMID: 16998613.

[22] P. M. Preshaw et al. “Periodontitis and diabetes: a two-way relationship.” In: *Diabetologia* 55.1 (2012), pp. 21–31. DOI: 10.1007/s00125-011-2342-y. PMID: 22057194.

[23] C.-W. J. Wang and L. K. McCauley. “Osteoporosis and Periodontitis.” In: *Current osteoporosis reports* 14.6 (2016), pp. 284–291. DOI: 10.1007/s11914-016-0330-3. PMID: 27696284.

[24] J. D. Beck et al. “Periodontal Disease and Coronary Heart Disease”. In: *Circulation* 112.1 (2005), 19 LP –24.

[25] H. Tilg and A. R. Moschen. “Adipocytokines: mediators linking adipose tissue, inflammation and immunity”. In: *Nature Reviews Immunology* 6.10 (2006), pp. 772–783. DOI: 10.1038/nri1937. PMID: 16998510.

[26] O. Osborn and J. M. Olefsky. "The cellular and signaling networks linking the immune system and metabolism in disease". In: *Nature Medicine* 18.3 (2012), pp. 363–374. DOI: 10.1038/nm.2627. PMID: 22395709.

[27] G. S. Hotamisligil et al. "Protection from obesity-induced insulin resistance in mice lacking TNF-alpha function." In: *Nature* 389.6651 (1997), pp. 610–614. DOI: 10.1038/39335. PMID: 9335502.

[28] R. Assuma et al. "IL-1 and TNF antagonists inhibit the inflammatory response and bone loss in experimental periodontitis." In: *Journal of immunology (Baltimore, Md. : 1950)* 160.1 (1998), pp. 403–9. PMID: 9551997.

[29] A. J. Delima et al. "Inflammation and Tissue Loss Caused by Periodontal Pathogens Is Reduced by Interleukin-1 Antagonists". In: *The Journal of Infectious Diseases* 186.4 (2002), pp. 511–516. DOI: 10.1086/341778. PMID: 12195378.

[30] D. Graves. "Cytokines That Promote Periodontal Tissue Destruction". In: *Journal of Periodontology* 79.8s (2008), pp. 1585–1591. DOI: 10.1902/jop.2008.080183. PMID: 18673014.

[31] E. Lalla, I. B. Lamster, and A. M. Schmidt. "Enhanced Interaction of Advanced Glycation End Products With Their Cellular Receptor RAGE: Implications for the Pathogenesis of Accelerated Periodontal Disease in Diabetes". In: *Annals of Periodontology* 3.1 (1998), pp. 13–19. DOI: 10.1902/annals.1998.3.1.13. PMID: 9722686.

[32] S. G. Grossi and R. J. Genco. "Periodontal Disease and Diabetes Mellitus: A Two-Way Relationship". In: *Annals of Periodontology* 3.1 (1998), pp. 51–61. DOI: 10.1902/annals.1998.3.1.51. PMID: 9722690.

[33] E. Lalla et al. "Hyperglycemia, glycoxidation and receptor for advanced glycation endproducts: potential mechanisms underlying diabetic complications, including diabetes-associated periodontitis." In: *Periodontology 2000* 23 (2000), pp. 50–62. PMID: 11276765.

[34] W. Zhao et al. "Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease". In: *Nature Genetics* 49.10 (2017), pp. 1450–1457. DOI: 10.1038/ng.3943. PMID: 28869590.

[35] NIH Consensus Development Panel on Osteoporosis Prevention, Diagnosis and Therapy. "Osteoporosis prevention, diagnosis, and therapy." In: *JAMA* 285.6 (2001), pp. 785–95. PMID: 11176917.

[36] N. C. Wright et al. "The Recent Prevalence of Osteoporosis and Low Bone Mass in the United States Based on Bone Mineral Density at the Femoral Neck or Lumbar Spine". In: *Journal of Bone and Mineral Research* 29.11 (2014), pp. 2520–2526. DOI: 10.1002/jbmr.2269.

[37] C. Weber and H. Noels. "Atherosclerosis: current pathogenesis and therapeutic options". In: *Nature Medicine* 17.11 (2011), pp. 1410–1422. DOI: 10.1038/nm.2538. PMID: 22064431.

[38] G. K. Hansson. "Inflammation, Atherosclerosis, and Coronary Artery Disease". In: *New England Journal of Medicine* 352.16 (2005), pp. 1685–1695. DOI: 10.1056/NEJMra043430.

[39] P. B. Lockhart et al. "Periodontal Disease and Atherosclerotic Vascular Disease: Does the Evidence Support an Independent Association?" In: *Circulation* 125.20 (2012).

[40] D. Shungin et al. "Using genetics to test the causal relationship of total adiposity and periodontitis: Mendelian randomization analyses in the Gene-Lifestyle Interactions and Dental Endpoints (GLIDE) Consortium". In: *International Journal of Epidemiology* 44.2 (2015), pp. 638–650. DOI: 10.1093/ije/dyv075. PMID: 26050256.

[41] J. Chhibber-Goel et al. "Linkages between oral commensal bacteria and atherosclerotic plaques in coronary artery disease patients." In: *NPJ biofilms and microbiomes* 2 (2016), p. 7. DOI: 10.1038/s41522-016-0009-7. PMID: 28649401.

[42] S. J. Leishman, H. Lien Do, and P. J. Ford. "Cardiovascular disease and the role of oral bacteria". In: *Journal of Oral Microbiology* 2.1 (2010), p. 5781. DOI: 10.3402/jom.v2i0.5781. PMID: 21523220.

[43] J. Erdmann et al. "A decade of genome-wide association studies for coronary artery disease: the challenges ahead". In: *Cardiovascular Research* (2018). DOI: 10.1093/cvr/cvy084. PMID: 29617720.

[44] A. S. Schaefer et al. "A genome-wide association study identifies GLT6D1 as a susceptibility locus for periodontitis." In: *Human molecular genetics* 19.3 (2010), pp. 553–62. DOI: 10.1093/hmg/ddp508. PMID: 19897590.

[45] N. T. Hashim et al. "Replication of the association of GLT6D1 with aggressive periodontitis in a Sudanese population". In: *Journal of Clinical Periodontology* 42.4 (2015), pp. 319–324. DOI: 10.1111/jcpe.12375.

[46] P. Deloukas et al. "Large-scale association analysis identifies new risk loci for coronary artery disease". In: *Nature Genetics* 45.1 (2012), pp. 25–33. DOI: 10.1038/ng.2480.

[47] A. S. Schaefer et al. "Genetic Evidence for PLASMINOGEN as a Shared Genetic Risk Factor of Coronary Artery Disease and Periodontitis". In: *Circulation: Cardiovascular Genetics* 8.1 (2015), pp. 159–167. DOI: 10.1161/CIRCGENETICS.114.000554. PMID: 25466412.

[48] T. W. T. C. C. Wellcome Trust Case Control Consortium. "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." In: *Nature* 447.7145 (2007), pp. 661–78. DOI: 10.1038/nature05911. PMID: 17554300.

[49] A. Helgadottir et al. "A Common Variant on Chromosome 9p21 Affects the Risk of Myocardial Infarction". In: *Science* 316.5830 (2007), pp. 1491–1493. DOI: 10.1126/science.1142842. PMID: 17478679.

[50] R. McPherson et al. "A common allele on chromosome 9 associated with coronary heart disease." In: *Science (New York, N.Y.)* 316.5830 (2007), pp. 1488–91. DOI: 10.1126/science.1142447. PMID: 17478681.

[51] A. S. Schaefer et al. "Identification of a shared genetic susceptibility locus for coronary heart disease and periodontitis." In: *PLoS genetics* 5.2 (2009), e1000378. DOI: 10.1371/journal.pgen.1000378. PMID: 19214202.

[52] A. P. Morris et al. "Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes". In: *Nature Genetics* 44.9 (2012), pp. 981–990. DOI: 10.1038/ng.2383. PMID: 22885922.

[53] G. Bochenek et al. "The large non-coding RNA ANRIL, which is associated with atherosclerosis, periodontitis and several forms of cancer, regulates ADIPOR1, VAMP3 and C11ORF10". In: *Human Molecular Genetics* 22.22 (2013), pp. 4516–4527. DOI: 10.1093/hmg/ddt299. PMID: 23813974.

[54] K. Divaris et al. "Genome-wide association study of periodontal pathogen colonization." In: *Journal of dental research* 91.7 Suppl (2012), 21S–28S. DOI: 10.1177/0022034512447951. PMID: 22699663.

[55] M. Munz et al. "A haplotype block downstream of plasminogen is associated with chronic and aggressive periodontitis". In: *Journal of Clinical Periodontology* (2017). DOI: 10.1111/jcpe.12749. PMID: 28548211.

[56] A. Teumer et al. "Genome-wide association study of chronic periodontitis in a general German population". In: *Journal of Clinical Periodontology* 40.11 (2013), pp. 977–985. DOI: 10.1111/jcpe.12154. PMID: 24024966.

[57] J. R. Shaffer et al. "Genome-wide association study of periodontal health measured by probing depth in adults ages 18-49 years." In: *G3 (Bethesda, Md.)* 4.2 (2014), pp. 307–14. DOI: 10.1534/g3.113.008755. PMID: 24347629.

[58] S. Freitag-Wolf et al. "Genome-wide exploration identifies sex-specific genetic effects of alleles upstream NPY to increase the risk of severe periodontitis in men". In: *Journal of Clinical Periodontology* 41.12 (2014), pp. 1115–1121. DOI: 10.1111/jcpe.12317. PMID: 25256105.

[59] P. Feng et al. "Genome wide association scan for chronic periodontitis implicates novel locus." In: *BMC oral health* 14 (2014), p. 84. DOI: 10.1186/1472-6831-14-84. PMID: 25008200.

[60] K.-W. Hong et al. "Genomewide association study on chronic periodontitis in Korean population: results from the Yangpyeong health cohort". In: *Journal of Clinical Periodontology* 42.8 (2015), pp. 703–710. DOI: 10.1111/jcpe.12437. PMID: 26212799.

[61] S. Shimizu et al. "A Genome-wide Association Study of Periodontitis in a Japanese Population". In: *Journal of Dental Research* 94.4 (2015), pp. 555–561. DOI: 10.1177/0022034515570315. PMID: 25672891.

[62] E. Kasbohm et al. "Exome Variant Analysis of Chronic Periodontitis in 2 Large Cohort Studies". In: *Journal of Dental Research* 96.1 (2017), pp. 73–80. DOI: 10.1177/0022034516665076. PMID: 27655622.

[63] A. Sanders et al. "Chronic Periodontitis Genome-wide Association Study in the Hispanic Community Health Study / Study of Latinos". In: *Journal of Dental Research* 96.1 (2017), pp. 64–72. DOI: 10.1177/0022034516664509. PMID: 27601451.

[64] M. I. McCarthy et al. "Genome-wide association studies for complex traits: consensus, uncertainty and challenges". In: *Nature Reviews Genetics* 9.5 (2008), pp. 356–369. DOI: 10.1038/nrg2344. PMID: 18398418.

[65] A. L. Price, C. C. A. Spencer, and P. Donnelly. "Progress and promise in understanding the genetic basis of common diseases". In: *Proceedings of the*

Royal Society B: Biological Sciences 282.1821 (2015), p. 20151684. DOI: 10.1098/rspb.2015.1684. PMID: 26702037.

[66] P. M. Visscher et al. "10 Years of GWAS Discovery: Biology, Function, and Translation". In: *The American Journal of Human Genetics* 101.1 (2017), pp. 5–22. DOI: 10.1016/j.ajhg.2017.06.005. PMID: 28686856.

[67] J. C. Barrett and L. R. Cardon. "Evaluating coverage of genome-wide association studies". In: *Nature Genetics* 38.6 (2006), pp. 659–662. DOI: 10.1038/ng1801. PMID: 16715099.

[68] W. S. Bush and J. H. Moore. "Chapter 11: Genome-wide association studies." In: *PLoS computational biology* 8.12 (2012), e1002822. DOI: 10.1371/journal.pcbi.1002822. PMID: 23300413.

[69] C. J. Hoggart et al. "Genome-wide significance for dense SNP and resequencing data". In: *Genetic Epidemiology* 32.2 (2008), pp. 179–185. DOI: 10.1002/gepi.20292. PMID: 18200594.

[70] the CARDIoGRAMplusC4D Consortium. "A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease." In: *Nature genetics* 47.10 (2015), pp. 1121–1130. DOI: 10.1038/ng.3396. PMID: 26343387.

[71] D. Ellinghaus et al. "Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci". In: *Nature Genetics* 48.5 (2016), pp. 510–518. DOI: 10.1038/ng.3528.

[72] N. Risch and K. Merikangas. "The future of genetic studies of complex human diseases." In: *Science (New York, N.Y.)* 273.5281 (1996), pp. 1516–7. PMID: 8801636.

[73] S. Zöllner and J. K. Pritchard. "Overcoming the Winner's Curse: Estimating Penetrance Parameters from Case-Control Data". In: *The American Journal of Human Genetics* 80.4 (2007), pp. 605–615. DOI: 10.1086/512821. PMID: 17357068.

[74] C. C. Chang et al. "Second-generation PLINK: rising to the challenge of larger and richer datasets". In: *GigaScience* 4.1 (2015), p. 7. DOI: 10.1186/s13742-015-0047-8. PMID: 25722852.

[75] Y. S. Aulchenko et al. "GenABEL: an R library for genome-wide association analysis." In: *Bioinformatics (Oxford, England)* 23.10 (2007), pp. 1294–6. DOI: 10.1093/bioinformatics/btm108. PMID: 17384015.

[76] B. Devlin and K. Roeder. "Genomic control for association studies." In: *Biometrics* 55.4 (1999), pp. 997–1004. DOI: 10.1111/j.0006-341X.1999.00997.x. PMID: 11315092.

[77] L. A. Hindorff et al. "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits". In: *Proceedings of the National Academy of Sciences* 106.23 (2009), pp. 9362–9367. DOI: 10.1073/pnas.0903103106. PMID: 19474294.

[78] J. MacArthur et al. "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)". In: *Nucleic Acids Research* 45.D1 (2017), pp. D896–D901. DOI: 10.1093/nar/gkw1133. PMID: 27899670.

[79] M. R. Nelson et al. "The support of human genetic evidence for approved drug indications". In: *Nature Genetics* 47.8 (2015), pp. 856–860. DOI: 10.1038/ng.3314. PMID: 26121088.

[80] E. Evangelou and J. P. A. Ioannidis. "Meta-analysis methods for genome-wide association studies and beyond". In: *Nature Reviews Genetics* 14.6 (2013), pp. 379–389. DOI: 10.1038/nrg3472.

[81] J. Higgins et al. "Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice". In: *Journal of Health Services Research & Policy* 7.1 (2002), pp. 51–61. DOI: 10.1258/1355819021927674. PMID: 11822262.

[82] J. Marchini and B. Howie. "Genotype imputation for genome-wide association studies". In: *Nature Reviews Genetics* 11.7 (2010), pp. 499–511. DOI: 10.1038/nrg2796.

[83] N. R. Roshyara et al. "Comparing performance of modern genotype imputation methods in different ethnicities". In: *Scientific Reports* 6.1 (2016), p. 34386. DOI: 10.1038/srep34386.

[84] B. Howie et al. "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing." In: *Nature genetics* 44.8 (2012), pp. 955–9. DOI: 10.1038/ng.2354. PMID: 22820512.

[85] Y. Li et al. "MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes". In: *Genetic Epidemiology* 34.8 (2010), pp. 816–834. DOI: 10.1002/gepi.20533. PMID: 21058334.

[86] C. Fuchsberger, G. R. Abecasis, and D. A. Hinds. "minimac2: faster genotype imputation". In: *Bioinformatics* 31.5 (2015), pp. 782–784. DOI: 10.1093/bioinformatics/btu704. PMID: 25338720.

[87] O. Delaneau, J. Marchini, and J.-F. Zagury. "A linear complexity phasing method for thousands of genomes." In: *Nature methods* 9.2 (2012), pp. 179–81. DOI: 10.1038/nmeth.1785. PMID: 22138821.

[88] B. N. Howie, P. Donnelly, and J. Marchini. "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies". In: *PLoS Genetics* 5.6 (2009). Ed. by N. J. Schork, e1000529. DOI: 10.1371/journal.pgen.1000529. PMID: 19543373.

[89] T. I. H. International HapMap Consortium. "A haplotype map of the human genome." In: *Nature* 437.7063 (2005), pp. 1299–320. DOI: 10.1038/nature04226. PMID: 16255080.

[90] R. M. Durbin et al. "A map of human genome variation from population-scale sequencing". In: *Nature* 467.7319 (2010), pp. 1061–1073. DOI: 10.1038/nature09534.

[91] T. 1. G. P. Consortium. "An integrated map of genetic variation from 1,092 human genomes". In: *Nature* 491.7422 (2012), pp. 56–65. DOI: 10.1038/nature11632.

[92] A. Auton et al. "A global reference for human genetic variation". In: *Nature* 526.7571 (2015), pp. 68–74. DOI: 10.1038/nature15393.

[93] S. McCarthy et al. "A reference panel of 64,976 haplotypes for genotype imputation." In: *Nature genetics* 48.10 (2016), pp. 1279–1283. DOI: 10.1038/ng.3643. PMID: 27548312.

[94] S. Das et al. “Next-generation genotype imputation service and methods”. In: *Nature Genetics* 48.10 (2016), pp. 1284–1287. DOI: 10.1038/ng.3656. PMID: 27571263.

[95] R. C. Lewontin and K.-i. Kojima. “THE EVOLUTIONARY DYNAMICS OF COMPLEX POLYMORPHISMS”. In: *Evolution* 14.4 (1960), pp. 458–472. DOI: 10.1111/j.1558-5646.1960.tb03113.x.

[96] P. Zavattari et al. “Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography, chromosome recombination frequency and selection”. In: *Human Molecular Genetics* 9.20 (2000), pp. 2947–2957. DOI: 10.1093/hmg/9.20.2947.

[97] M. W. Nachman. “Variation in recombination rate across the genome: evidence and implications”. In: *Current Opinion in Genetics & Development* 12.6 (2002), pp. 657–663. DOI: 10.1016/S0959-437X(02)00358-1.

[98] R. C. Lewontin. “THE INTERACTION OF SELECTION AND LINKAGE. I. GENERAL CONSIDERATIONS; HETEROTIC MODELS”. In: *Genetics* 49.1 (1964).

[99] F. W. Albert and L. Kruglyak. “The role of regulatory variation in complex traits and disease”. In: *Nature Reviews Genetics* 16.4 (2015), pp. 197–212. DOI: 10.1038/nrg3891. PMID: 25707927.

[100] G. GTEx Consortium. “Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.” In: *Science (New York, N.Y.)* 348.6235 (2015), pp. 648–60. DOI: 10.1126/science.1262110. PMID: 25954001.

[101] J. D. Eicher et al. “GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes.” In: *Nucleic acids research* 43.Database issue (2015), pp. D799–804. DOI: 10.1093/nar/gku1202. PMID: 25428361.

[102] T. Lappalainen et al. “Transcriptome and genome sequencing uncovers functional variation in humans.” In: *Nature* 501.7468 (2013), pp. 506–11. DOI: 10.1038/nature12531. PMID: 24037378.

[103] E. R. Gamazon et al. “SCAN: SNP and copy number annotation.” In: *Bioinformatics (Oxford, England)* 26.2 (2010), pp. 259–62. DOI: 10.1093/bioinformatics/btp644. PMID: 19933162.

[104] K. Xia et al. “seeQTL: a searchable database for human eQTLs.” In: *Bioinformatics (Oxford, England)* 28.3 (2012), pp. 451–2. DOI: 10.1093/bioinformatics/btr678. PMID: 22171328.

[105] H.-J. Westra et al. “Systematic identification of trans eQTLs as putative drivers of known disease associations.” In: *Nature genetics* 45.10 (2013), pp. 1238–43. DOI: 10.1038/ng.2756. PMID: 24013639.

[106] C.-H. Yu, L. R. Pal, and J. Moult. “Consensus Genome-Wide Expression Quantitative Trait Loci and Their Relationship with Human Complex Trait Disease.” In: *Omics : a journal of integrative biology* 20.7 (2016), pp. 400–14. DOI: 10.1089/omi.2016.0063. PMID: 27428252.

[107] A. Ramasamy et al. “Genetic variability in the regulation of gene expression in ten regions of the human brain.” In: *Nature neuroscience* 17.10 (2014), pp. 1418–1428. DOI: 10.1038/nn.3801. PMID: 25174004.

[108] K. Suhre et al. "Connecting genetic risk to disease end points through the human blood plasma proteome." In: *Nature communications* 8 (2017), p. 14357. DOI: 10.1038/ncomms14357. PMID: 28240269.

[109] E. Lieberman-Aiden et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." In: *Science (New York, N.Y.)* 326.5950 (2009), pp. 289–93. DOI: 10.1126/science.1181369. PMID: 19815776.

[110] J. R. Dixon et al. "Topological domains in mammalian genomes identified by analysis of chromatin interactions." In: *Nature* 485.7398 (2012), pp. 376–80. DOI: 10.1038/nature11082. PMID: 22495300.

[111] M. Kircher et al. "A general framework for estimating the relative pathogenicity of human genetic variants". In: *Nature Genetics* 46.3 (2014), pp. 310–315. DOI: 10.1038/ng.2892. PMID: 24487276.

[112] L. N. Borrell and P. N. Papapanou. "Analytical epidemiology of periodontitis". In: *Journal of Clinical Periodontology* 32.s6 (2005), pp. 132–158. DOI: 10.1111/j.1600-051X.2005.00799.x. PMID: 16128835.

[113] R. D. Vaithilingam et al. "Moving into a new era of periodontal genetic studies: relevance of large case-control samples using severe phenotypes for genome-wide association studies". In: *Journal of Periodontal Research* 49.6 (2014), pp. 683–695. DOI: 10.1111/jre.12167. PMID: 24528298.

[114] A. K. Grotz, A. L. Gloyn, and S. K. Thomsen. "Prioritising Causal Genes at Type 2 Diabetes Risk Loci". In: *Current Diabetes Reports* 17.9 (2017), p. 76. DOI: 10.1007/s11892-017-0907-y. PMID: 28758174.

[115] S. Smemo et al. "Obesity-associated variants within FTO form long-range functional connections with IRX3". In: *Nature* 507.7492 (2014), pp. 371–375. DOI: 10.1038/nature13138. PMID: 24646999.

[116] S. T. Nurnberg et al. "From Loci to Biology: Functional Genomics of Genome-Wide Association for Coronary Disease". In: *Circulation Research* 118.4 (2016), pp. 586–606. DOI: 10.1161/CIRCRESAHA.115.306464. PMID: 26892960.

[117] E. Evangelou et al. "Impact of Phenotype Definition on Genome-Wide Association Signals: Empirical Evaluation in Human Immunodeficiency Virus Type 1 Infection". In: *American Journal of Epidemiology* 173.11 (2011), pp. 1336–1342. DOI: 10.1093/aje/kwr024.

[118] B. K. Bulik-Sullivan et al. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies". In: *Nature Genetics* 47.3 (2015), pp. 291–295. DOI: 10.1038/ng.3211.

[119] B. Bulik-Sullivan et al. "An atlas of genetic correlations across human diseases and traits". In: *Nature Genetics* 47.11 (2015), pp. 1236–1241. DOI: 10.1038/ng.3406.

[120] O. Zuk et al. "The mystery of missing heritability: Genetic interactions create phantom heritability." In: *Proceedings of the National Academy of Sciences of the United States of America* 109.4 (2012), pp. 1193–8. DOI: 10.1073/pnas.1119675109. PMID: 22223662.

[121] T. R. Webb et al. "Systematic Evaluation of Pleiotropy Identifies 6 Further Loci Associated With Coronary Artery Disease". In: *Journal of the American*

College of Cardiology 69.7 (2017), pp. 823–836. DOI: 10.1016/j.jacc.2016.11.056. PMID: 28209224.

[122] E. A. Boyle et al. "An Expanded View of Complex Traits: From Polygenic to Omnigenic". In: *Cell* 169.7 (2017), pp. 1177–1186. DOI: 10.1016/j.cell.2017.05.038. PMID: 28622505.

[123] A. V. Khera et al. "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations". In: *Nature Genetics* (2018). DOI: 10.1038/s41588-018-0183-z. PMID: 30104762.

[124] N. R. Wray et al. "Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model." In: *Cell* 173.7 (2018), pp. 1573–1580. DOI: 10.1016/j.cell.2018.05.051. PMID: 29906445.

[125] G. Genovese et al. "Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia". In: *Nature Neuroscience* 19.11 (2016), pp. 1433–1441. DOI: 10.1038/nn.4402. PMID: 27694994.

[126] C. Fuchsberger et al. "The genetic architecture of type 2 diabetes". In: *Nature* 536.7614 (2016), pp. 41–47. DOI: 10.1038/nature18642. PMID: 27398621.

[127] Y. Luo et al. "Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7". In: *Nature Genetics* 49.2 (2017), pp. 186–192. DOI: 10.1038/ng.3761. PMID: 28067910.

[128] K. Zhong et al. "CollapsABEL: an R library for detecting compound heterozygote alleles in genome-wide association studies." In: *BMC bioinformatics* 17 (2016), p. 156. DOI: 10.1186/s12859-016-1006-9. PMID: 27059780.

Curriculum Vitae

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten.

List of Publications

List of my publications that were under review or published until December 2018. Equal, shared first authorships are marked with *.

Under review

Munz M et al. Qtlizer: comprehensive QTL annotation of GWAS results

Clark et al. Effects of autozygosity on a broad range of human phenotypes

Kessler et al. Association of the coronary artery disease risk gene GUCY1A3 with on-aspirin platelet reactivity and ischemic events after coronary intervention

2018

Munz M et al. (2018) Meta-analysis of genome-wide association studies of aggressive and chronic periodontitis identifies two novel risk loci. *Eur J Hum Genet.* PMID: 30218097

Munz M et al. (2018) Genome-wide association meta-analysis of coronary artery disease and periodontitis reveals a novel shared risk locus. *Sci Rep.* PMID: 30209331

Offenbacher et al. (2018) GWAS for Interleukin-1β levels in gingival crevicular fluid identifies IL37 variants in periodontal inflammation. *Nat Commun.* PMID: 30206230

Westermaier AL*, **Munz M*** et al. (2018) Association of genetic variation at AQP4 locus with vascular depression. *Biomolecules.* PMID: 30563176

Schaich A, Westermaier AL, **Munz M** et al. (2018) Mental health and psychosocial functioning over the lifespan of German patients undergoing cardiac catheterization for coronary artery disease. *Front Psychiatry.* PMID: 29593584

Aarabi G, Zeller T, Heydecke G, **Munz M**, Schaefer A, Seedorf U (2018) Roles of the Chr.9p21.3 ANRIL Locus in Regulating Inflammation and Implications for Anti-Inflammatory Drug Target Identification. *Front Cardiovasc Med.* PMID: 29868613

Nashef A, Qabaja R, Salaymeh Y, Botzman M, **Munz M**, et al. (2018) Integration of Murine and Human Studies for Mapping Periodontitis Susceptibility. *J Dent Res.* PMID: 29294296

2017

Munz M et al. (2017) A genome-wide association study identifies nucleotide variants at SIGLEC5 and DEFA1A3 as risk loci for periodontitis. *Hum Mol Genet.* PMID: 28449029

Munz M*, Chen H* et al. (2017) A haplotype block downstream of plasminogen is associated with chronic and aggressive periodontitis. *J Clin Periodontol.* PMID: 28548211

Shusterman A, **Munz M** et al. (2017) The PF4/PPBP/CXCL5 Gene Cluster Is Associated with Periodontitis. *J Dent Res.* PMID: 28467728

Eitel I, Moeller C, **Munz M** et al. (2017) Genome wide association study in takotsubo syndrome - Preliminary results and future directions. *Int J Cardiol.* PMID: 28108127

Haycock PC et al. (2017) Association Between Telomere Length and Risk of Cancer and Non-Neoplastic Diseases. *JAMA Oncol.* PMID: 28241208

2016

Offenbacher S et al. (2016) Genome-wide association study of biologically informed periodontal complex traits offers novel insights into the genetic basis of periodontal disease. *Hum Mol Genet.* PMID: 26962152

2015

Munz M, Tönnies S, Balke W-T, Simon E (2015) Multidimensional gene search with Genehopper. *Nucleic Acids Res.* PMID: 25990726

Tinhofer I, Niehr F, Konschak R, Liebs S, **Munz M**, Stenzinger A, Weichert W, Keilholz U, Budach V (2015) Next-generation sequencing: hype and hope for development of personalized radiation therapy? *Radiat Oncol.* PMID: 26316159

Declaration of Independence

I, Matthias Munz, declare the following statements as true: I have independently authored this thesis; I have not used other sources and/or resources used in this thesis; I have explicitly marked all material which has been quoted either literally or by content from the used sources; this thesis has not been submitted in this or any other form to other examination authority; and that the included publications were not part of another thesis using the same personal contributions.

Signature	Date