Aus dem Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie, Lehrstuhl Biometrie und Bioinformatik,
Ludwig-Maximilians-Universität München
Vorstand: Prof. Dr. Ulrich Mansmann

Mechanism-driven Hypothesis Generation Support for a Predictive Adverse Effect in Colorectal Cancer Treatment

Dissertation zur Erlangung des Doktorgrades der Naturwissenschaften an der Medizinischen Fakultät der Ludwig-Maximilians-Universität zu München

vorgelegt von
Sebastian Schaaf
aus
Bergisch Gladbach

2018

Mit Genehmigung der Medizinischen Fakultät
der Universität München

Betreuer:	Prof. Dr. Ulrich Mansmann
Zweitgutachter:	Prof. Dr. Volker Heun
Dekan:	Prof. Dr. med. dent. Reinhard Hickel
Tag der mündlichen Prüfung:	15.02.2019

Eidesstattliche Versicherung

Schaaf, Sebastian

Name, Vorname

Ich erkläre hiermit an Eides statt,

dass ich die vorliegende Dissertation mit dem Thema

„Mechanism-driven Hypothesis Generation Support for a Predictive Adverse Effect in Colorectal Cancer Treatment“

selbstständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

Kerpen, 15.01.2020 — Ort, Datum

Sebastian Schaaf — Unterschrift Doktorandin/Doktorand

Schule ist aus…

(für Opa)

Zusammenfassung

Diese bioinformatische Dissertation beschreibt die tumorbiologische Hypothesengenierung, insbesondere im Kontext des Kolorektalkarzinoms.

Hintergrund der Studien ist eine Beobachtung aus der klinischen Praxis. Verschiedene Autoren berichten, dass bei der Behandlung mit Inhibitoren des Epidermalen Wachstumsfaktor Rezeptors (EGFR), speziell des therapeutischen Antikörpers Cetuximab, eine Minderheit der Patienten die übliche Nebenwirkung der Hauttoxizität nicht oder in deutlich verminderter Form zeigt. Bei diesen Patienten wird gleichzeitig eine reduzierte Wirksamkeit der Therapie beschrieben. Das Ausbleiben der Nebenwirkung wird somit als phänotypischer Biomarker genutzt, um gegebenenfalls die Therapie anzupassen. Nachteilig erscheint in diesem Kontext allerdings die präventive Hautpflege sowie die Tatsache, dass eine Cetuximab-Behandlung zunächst gestartet werden muss, um eine Information über die Wirksamkeit zu gewinnen. Dadurch, dass der zugrunde liegende molekulare Mechanismus unbekannt ist, kann keine Vorhersage anhand eines klinischen Test getroffen werden.

In der vorliegenden Arbeit war es das Ziel, Hypothesen zu generieren, welche Proteine und zellulären Signalwege kausal für das unterschiedliche Ansprechverhalten der Patientengruppen sein könnten. Ausgehend von der Annahme, dass natürliche Keimbahnvarianten in der Erbinformation der Individuen im Behandlungskontext diskriminatorisch wirken, baut die Dissertation auf einem kleinen Datensatz von 23 Exomen von Teilnehmern klinischer Studien auf. Diese Sequenzierungsdaten wurden in genomische Varianten überführt und auf ihren potentiellen genetisch-mechanistischen Einfluss hin untersucht. Gezielte Einschränkungen wurden dabei anhand einer Modellierung des biomedizinischen Kontextes des Anwendungsfalls eingeführt, um die reduzierte Datenlage gezielt mit Informationen anzureichern. Die so erhaltenen Kandidatengene, welche in nachfolgenden praktischen Arbeiten validiert werden müssen, werden im Einzelnen beschrieben und bewertet.

Methodisch ist das Ergebnis dieser Dissertation die „Molecular Systems Map“, eine in Cytoscape modellierte Netzwerkstruktur, die funktionelle Interaktionen zwischen Proteinen interaktiv visualisiert und gleichzeitig als Filter auf Basis des biologischen Kontexts dient. Ziel hierbei ist es, einen biomedizinisch ausgebildeten Fachanwender bei der Generierung von Hypothesen zu unterstützen, indem im Gegensatz zu sonst häufig anzutreffenden tabellarischen Ansichten die Ergebnisse aus der Sequenzanalyse in eben jenem funktionalen Kontext dargestellt werden. Darüber hinaus wird so die Anwendung von Graphenalgorithmen und die Integration weiterer Daten ermöglicht, z.B. solcher aus komplementären ‘omics-Experimenten.

Abstract

This bioinformatics thesis describes work and results from a study on a use case in the context of colorectal cancer.

Background of the studies is an observation form the clinical practice. Various authors report that upon treatment with inhibitors of the Epidermal Growth Factor Receptor (EGFR), in particular with the therapeutic antibody Cetuximab, a minority of patients does not, or in a clearly reduced form, show common adverse effects of skin toxicity. For these patients, at the same time a reduced efficacy of the therapy is described. The lack of the adverse effect therefore gets used as a phenotypic biomarker for inducing a switch of therapy. However, preventive skin care during treatment, counteracting the biomarker signal, and the necessity to start the therapy first in order to gain the information, appear unfavorable. As the underlying molecular mechanisms remain elusive, predictions ahead of treatment, e.g. by a clinical test, are not possible yet.

In the presented work, the aim was to generate hypotheses, which proteins and cellular signaling pathways might be causal for the differentiating response of the patient groups. Starting from the assumption that naturally occurring germline variations functionally discriminate individuals in the context of the treatment, the thesis builds up on a small dataset of 23 exomes of patients from a clinical study context. These sequencing data were processed to genomic variants and analyzed for their potential influence on the mechanistic level. Targeted restrictions were introduced by modeling the biomedical context of the use case in order to enrich the sparse individual data with further information. The obtained candidate genes, which are necessary to be validated in practical studies, are described and evaluated in detail.

Methodologically, the result of the thesis is the „Molecular Systems Map“, a network data structure modeled in Cytoscape, interactively visualizing the functional interactions of proteins and simulatenously filtering the called variants upon the biological context. Here, the aim is to enable biomedical domain experts, beyond scrolling tabular information on called variants, to review their experimental data in the functional context and support them in the hypothesis generation process. Additionally, this provides the opportunity to apply graph algorithms and integrate further data, e.g. such from completary ‘omics experiments.

Contents

Figures

Tables

Boxes

1. Introduction

This thesis is set at the intersection of computer science, biology and medicine. From a clinical of view, the subject of interest is colorectal cancer, which for a relevant amount of patients is treated using therapeutic antibodies targetting the Epidermal Growth Factor Receptor (EGFR), being a key molecular factor promoting tumor growth. However, for a minority of patients, anti-EGFR treatment fails, indicated by missing skin toxicity reactions, which are a usual adverse effect. Meanwhile, the rationals for this observation, subsequently named “skin rash phenomenon”, remains elusive. In terms of tumor biology, this raises the question for the underlying molecular mechanisms. Although in cancer settings plenty of knowledge has been collected over the years, a transparent model allowing prediction of treatment success upon any measurement does not yet exist. Therefore, this thesis targets a small number of patient-derived next-generation sequencing datasets and aims to bring the individual variations into a general mechanistic context derived from literature. For the latter, a number of considerations were made on potentially involved proteins and mechanisms from cellular signaling, immunology and cell adhesion.

Taken together, the following sections provide deeper insights into the factors building grounds for this thesis. Starting clinically from CRC and its treatment (Section 1.1), tumor biology with a focus on EGFR signaling is highlighted (Section 1.2). Section 1.3 is based to a larger extend on a review on the skin rash phenomenon by Mario Lacouture [Lacouture 2006] and related work. These publications in particular, and a range of more general knowledge, formed ground for the basic assumptions and rationals (Section 1.4). Finally, as the integrative target data structure and user interface has been decided to be a visual network, a short corresponding review and argumentation is given in Section 1.5.

1.1 Epidemiological and clinical aspects of colorectal cancer

Colorectal cancer (CRC) as third most diagnosed cancer in both women and men [Siegel *et al.* 2012] is an epidemiologically important factor among cancer cases worldwide, accounting for >1.2 Mio. diagnoses per year, 520,000 of them registered in the western hemisphere [Ferlay *et al.* 2010]. In Germany, every seventh diagnosed cancer is CRC. Mortality rates are decreasing, ranging from 35% to 50% [Moriarity *et al.* 2016]. With 26,000 deaths per year in Germany due to CRC, absolute numbers are still high, meanwhile, 478,000 new cases were detected [Robert Koch-Institut 2015]. Approximately, every second patient is diagnosed to already suffer from metastatic CRC (mCRC), lowering chances for recurrence-free curation [Tjandra & Chan 2007].

Considering these numbers and the fact that in both the US and Germany every fourth death is caused by cancer [Siegel *et al.* 2012; Robert Koch-Institut 2015], effective treatment of CRC is a crucial factor in health care.

CRC typically originates from benign tumors forming polyps. Symptoms like longer lasting diarrhea, cramping pain, rectal bleeding etc. occur late in tumor development and are ambiguous for other diseases and intolerances, too. Consequently, definite diagnoses are often made in late stages, in which the cancer has already grown into the surrounding tissue and even metastasized to other organs [NIH-NCI 2016a].

Following the American Cancer Society's guidelines [NIH-NCI 2016b], standard of care for both stages 0 (no growth of the tumor through the colon's inner lining) or I (penetration, but no growth beyond the colon wall or nearby lymph nodes) is resection of the polyp and, if necessary, parts of the colon. While for stage II (spreading to nearby tissues, but not lymph nodes), adjuvant chemotherapy after surgery may be considered in order to avoid recurrences, it is the standard treatment for stage III CRC. In this stage, nearby lymph nodes are affected, but no other organs. Potentially detached cancer cells may be targetted by radiation therapy.

For stage IV CRC, cancer cells have metastasized into other organs and tissues most likely to liver, lungs, the peritoneum or distant lymph nodes. In this systemic setting, surgery is regarded as unlikely to be useful. Neoadjuvant chemotherapy may be a treatment option to weaken and shrink cancerous sites first, improving chances for curative surgery. Apart from selected chemotherapeutics and often combined like 5-Fluorouracil (5-FU), oxaliplatin, irinotecan and other cytostatic drugs, targeted therapies invoking monoclonal antibodies (mAbs) have entered the clinical practice. Recently, a notable positive effect of these new biologic drugs has been described, as quality of life (QoL), 5-year and overall survival have been significantly improved for advanced CRC [Wen & Li 2016]. The combination therapies are subject to recent clinical studies on effectiveness in the context of the stage, molecular status and even location of the tumor (comp. e.g. [Ciliberto *et al.* 2018]). Additionally, decisions for or against a certain medication depend on the patient's health parameters as well as on the personal genomic background [Ab Mutalib *et al.* 2017; Patel *et al.* 2018]. Anyhow, although pharmacogenetics/-genomics have been described as a hope for individualized medicine some years ago (comp. [Ma & Lu 2011]), practically applied knowledge appears sparse.

Nowadays, the signaling pathways of two growth receptors are primarily tackled, namely these of Epidermal Growth Factors (EGF and similar) and the Vascular Endothelial Growth Factor (VEGF). Both are classical representatives for two of the popular 'cancer hallmarks' [Hanahan & Weinberg 2000, 2011]. In the case of a de-regulated EGF receptor (EGFR) pathway, primarely proliferative signaling is

sustained, enabling the hallmark 'self-sufficiency in growth signals', a basic feature of tumorous growth. However, the EGFR pathway also stimulates another five of the six cancer hallmarks of the first model of Hanahan and Weinberg: independence of growth signals, insensitivity to growth-inhibitory signals, resistance to programmed cell death, angiogenesis, and metastasis [Gazdar 2009].

VEGF induces and sustains the more advanced hallmark of angiogenesis, which describes the formation of local blood vessels through vasculogenesis (*de novo*) and angiogenesis (from pre-existing vessels). In the context of cancer, tumors of a size beyond 1-2 mm require dedicated ingrowth of capillaries, as the process of diffusion of oxygen and nutrients through tissues gets insufficient for further progress of the malignant cell mass [McDougall *et al.* 2006]. Gaining a mutation that enables a tumor to overexpress VEGF therefore is denoted as 'angiogenic switch' and marks an important point in tumor development. In turn, therapeutic shutdown of VEGF signaling (e.g. ligand blocking using the therapeutic antibody Bevacizumab [Ohhara *et al.* 2016]) is intended to inhibit the formation of new vessels, to abolish immature ones and to decrease the vascular wall's perfusion rates. Consequently, the blood-derived supply with oxygen and nutrients is impaired, tumors which already facilitated the angiogenic switch may shrink [Willett *et al.* 2004].

1.2 EGFR signaling and Cetuximab

The epidermal growth factor receptor (EGFR), discovered 1978, is a transmembrane receptor tyrosine kinase and a prototypic representative for both the ErbB receptor family as well as the superfamily of receptor tyrosine kinases (RTKs) [Carpenter *et al.* 1978; Wells 1999; Seshacharyulu *et al.* 2012]. RTKs are cell surface receptors, which show high affinity to growth factors, cytokines and hormones [Robinson *et al.* 2000], and encompass also the VEGF receptor (VEGFR). In general, those molecular on/off switches forward signals from the cell's surface to downstream cascades by phosphorylation of the intracellular thyrosine kinase domain, resulting in variation of gene expression.

The ErbB family consists of four known members in humans (ErbB1/EGFR/HER1, ErbB2/HER2/Neu, ErbB3/HER3 and ErbB4/HER4) [Seshacharyulu *et al.* 2012], of which the first two are known to be overactive in various solid tumors [Cho & Leahy 2002]. For EGFR, at least seven different activating ligands are known, including EGF, transforming growth factor-alpha (TGFα) and heparin-binding EGF (HB-EGF) [Linggi & Carpenter 2006]. These trigger various dimerizations of ErbB family member monomers [Schlessinger 2002], reflecting the general ability of RTKs to expand both the spectrum of bindable ligands and the diversity of recruited downstream pathways [Holland *et al.* 2003].

Cellular target mechanisms of EGFR signaling are as diverse as important. They include protein secretion, cell motility, the balance between mitogenesis or apoptosis as well as differentiation and dedifferentiation. In cancer settings, the ability of the EGFR pathway to promote survival and growth, is an important factor for tumor progression, tissue invasion and metastasis [Wells 1999].

Although the reported percentage of tumors (CRC, but also lung, head and neck, and gliomas) varies between 25% and 82% [Krasinskas 2011], a majority of cancer cases is reported to overexpress EGFR, showing an association with progression to malignancy, more advanced tumor stages, poor differentation, worse histologic grade, lymph vascular invasion and finally, poor prognosis [Yarom & Jonker 2011; Grandis & Sok 2004]. EGFR as a target for cancer therapy was already focussed in the 1980s [Aboud-Pirak *et al.* 1988], resulting in the first therapeutics fifteen years later. All those inhibitory drugs belong to either the class of tyrosine kinase inhibitors (TKIs) or monoclonal antibodies (mAbs) [Melosky *et al.* 2009]. TKIs in general are small molecules, acting on the intracellular kinase domain, blocking downstream signaling. In contrast, therapeutic antibodies act on the extracellular receptor domain. Additional to blocking EGFR signaling, it is assumed that immune responses against the mAB's tail, e.g. via the complement system, may also contribute an effect [Holubec *et al.* 2016].

Gefitinib was the first TKI, approved for the treatment in non-small cell lung cancer (NSCLC) since 2002 in Japan [evaluate.com 2002], 2003 in the US [NIH-NCI 2011] and 2009 in Europe [EMA 2009], respectively. Erlotinib, another TKI, received approval in 2005. In 2004, with Cetuximab the first antibody directed against EGFR entered the market [NIH-NCI 2013a], followed by Panitumumab in 2006 [NIH-NCI 2013b]. Both were dedicated to CRC first [Ohhara *et al.* 2016]. For both TKIs and antibodies, further molecules became available and for most of them the range of application was iteratively extended to other cancer entities with overexpression of EGF(R).

Shortly after those new drugs had entered the market and clinical practice, conditionals in treatment emerged. On the one hand, genetic factors like target variations in the tumor (e.g. EGFR mutations [Paez *et al.* 2004; Lynch *et al.* 2004]) or ethnical background (e.g. Asian or European origin), but also further epidemiologic parameters like environmental influences (smoker or not) show an effect on treament efficacy [Gazdar 2009; Stuart & Sellers 2009]. On the other hand, the treatment procedures, especially the combination and sequence of options (radiation, various chemotherapeutics, TKIs, mAbs), are subject of intensive discussions and clinical studies. While (neo-)adjuvant chemotherapy plus blockade of EGFR are intended to be more effective than monotherapy [Yazdi *et al.* 2015], other authors decline any effect in adjuvant settings and a controversial one in neo-adjuvant treatments, where the aim is to shrink the tumor in order to allow resection [Schmiegel *et al.* 2008]. Also common mutations in functionally connected genes are accepted to be crucial for certain decisions – for

example, activating KRAS mutations exclude the application of EGFR inhibitors, as proliferation signals are generated downstream of EGFR in this setting [Karapetis *et al.* 2008].

While KRAS mutations are known to be the causal reason for a lack of therapeutic effect, mechanistic backgrounds of side effects remain elusive. Consequently, the occurrence of skin rash as an adverse effect of Cetuximab application can be used as a predictive biomarker for treatment efficacy, but not predicted itself.

1.3 Cetuximab, skin rash and anti-tumor efficacy

Even before FDA's approval of EGFRIs, their dermatologic side effects were reported [Herbst *et al.* 2003] and became subject to various intensive discussions, dedicated to classification, treatment guidelines, the patient's psycho-social discomfort and the urgent need for mechanistic understanding.

Skin toxicity as an adverse effect in EGFRI treatment describes a variety of symptoms affecting the epidermis, hair follicles and periungual tissues [Lacouture 2006]. Usual symptoms are papulopustular rash in sun-exposed areas like head (scalp and eyelashes), chest (V-shaped area) and legs, dry and itchy skin, pruritus, abnormalities of hair and nails, up to pain in extreme cases [Robert *et al.* 2005]. The pathology overall has often been named as 'acne-like' by various authors, although there are no shared features in terms of histopathology and pharmacologic response (reaction to anti-inflammatory drugs, but not anti-acne agents; [Lacouture 2006]). By observation, there is a certain sequence of symptoms observed in the majority of patients (comp. Fig. 1).

In the affected areas, tissues are constantly growing, triggered by EGF signaling. The activation of EGFR leads to a controlled sequence of cell proliferation, migration and differentiation, primarely for keratinocytes [Jost *et al.* 2000]. Consequently, the skin shows a high degree of self-organization when viewed in cross section (comp. Fig. 2). Generally, the skin subdivides into multiple layers: hypodermis, dermis (both contain blood vessels), epidermis and finally the stratum corneum. The basement membrane, separating dermis and epidermis, is the location of undifferentiated, basal keratinocytes, which in turn make up the major epidermal cell type [Lacouture 2006]. Most importantly, developmental physiology of the skin depends on keratinocyte maturation, which includes a migratory movement towards the skin surface and ends in apoptosis [Candi *et al.* 2005]. The terminal differentiated cells are designated as corneocytes. Maturation as well as the fragile balance between migration and adhesion is highly regulated [Fuchs & Raghavan 2002]. Although the EGFR pathway orchestrates these mechanisms, the molecule itself is expressed in keratinocytes of the basal and suprabasal layers only [Nanney *et al.* 1990].

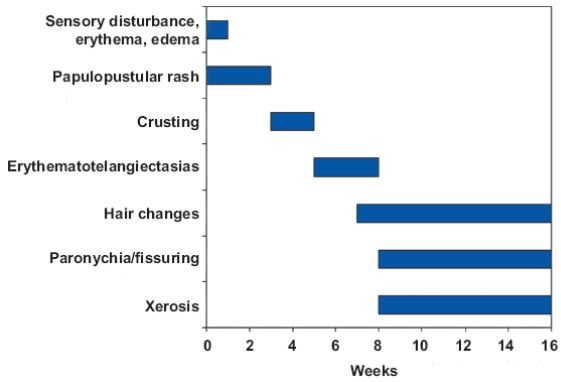


Fig. 1 - Sequence of rash events in EGFRI-treated patients. While skin-related effects occur immediately in the first week upon EGFRI application, disturbances of (slower growing) hair and nails take weeks to manifest. Xerosis, presumably through hypofunction of sebaceous glands, also appears as late symptom. In between, crusting and visibly dilated vessels in the skin appear (from: [LoRusso 2009]).

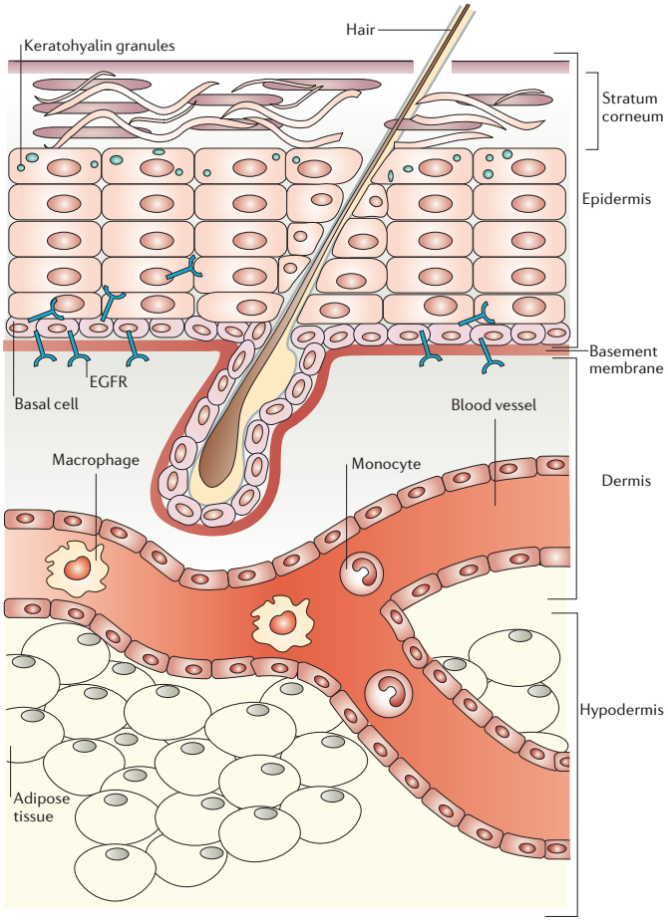


Fig. 2 - The structure of the skin and EGFR location. The multilayer model of the skin from Lacouture 2006. While below the basement membrane dermis and hypodermis are supplied by bloods vessels, the epidermis consists to approximately 90% of keratinocytes. At hair follicles, the basement membrane is invaginated, but continuous. Here, like everywhere immediately at the epidermal side of the basement membrane, undifferentiated keratinoycytes are constantly proliferated from stem cells. These basal keratinocytes express EGFR. On their way to the skin's surface, EGFR expression gets lost soon. However, activation of the EGFR pathway triggers the differentatiation of keratinocytes to terminal corneocytes. During the progression from epidermis to the stratum corneum, keratohyalin is massively expressed and stored in granules. Meanwhile, the cell undergoes apoptosis. The final anucleated corneocytes make up the protective layer against the outside.

Vice versa, inhibition of EGFR disrupts the skin's integrity by accelerating the maturation of keratinocytes by arresting growth and migration (comp. Fig. 3). Premature keratinocytes proceed to terminal differentiation too early, reducing the epidermal thickness and impairing the stratum corneum. Simultaneously, EGFR blockade triggers recruitment of immune cells, causing local inflammation, apoptosis, tissue damage and dilated vessels [Lacouture 2006]. While the skin's barrier function to the extracorporal environment is affected overall, simultaneously bacteria and other microbes can enter the skin, further amplifying the immune cell's reactions [Eilers *et al.* 2010].

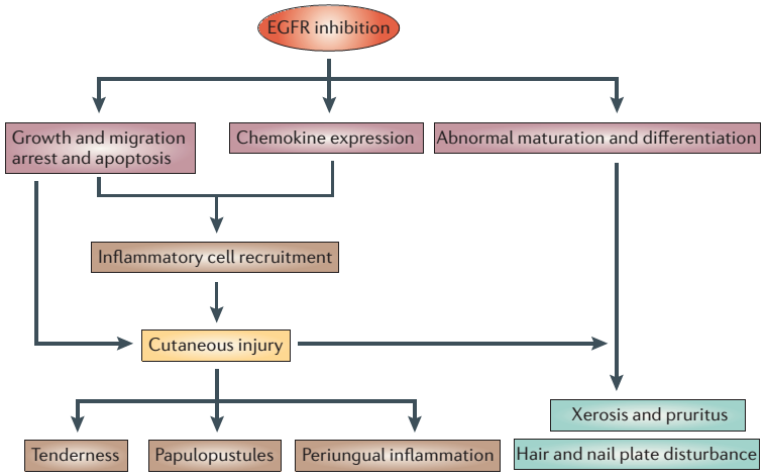


Fig. 3 - Flowchart model of EGFRI-induced reactions of the skin according to Lacouture 2006. Inhibition of EGFR leads to arrest of growth and migration, apoptosis as well as chemokine expression immediately. Consequently, inflammatory cells get locally recruited, amplifying cutaneous injuries. These include tenderness of the skin and rash, at hair follicles and nail. Longterm effects like xerosis, pruritus and macroscopic disturbances of hair and nails rely on persistent abnormal maturation and differentiation of the skin tissues.

Although for the majority of patients the EGFRI side effects are mild, <15 % develop severe toxicities [Melosky *et al.* 2009], which significantly lower the quality of life. In these cases, dose modifications (three quarters of cases), treatment delays or even discontinuation of therapy is the consequence for a third of the patients [Lacouture 2009].

Therefore, rash management is key during EGFRI treatment, supporting the patient tolerance and compliance [Lacouture & Melosky 2007]. Consequently, over the last decade dermatology-driven recommendations for preventative/prophylactic and reactive treatment of skin toxicities have been published (e.g. [Lacouture *et al.* 2011]). As most of the clinical study reports use the National Cancer Institute's Common Terminology Criteria for Adverse Effects (NCI CTCAE; current version is 5.0 [NIH-NCI 2017]), classifiers for reactive treatment start from a diagnosed grade of toxicity, ranging from mild (grade 1) over moderate (grade 2) up to severe (grade 3/4) adverse effects. Usual treatments are

topical application of moisturizers, suncream and hydrocortisone, systemic reactions are counteracted with Minocycline or Doxycycline orally.

However, the CTCAE to date is not primarely designed for classifying EGFRI-induced skin toxicities, leading to underreported severity and consequently underadjustment of therapy. In turn, the Multinational Association for Supportive Care in Cancer (MASCC) offer their EGFRI Skin Toxicity Tool (MESTT) [Lacouture *et al.* 2010]), correlating well with CTCAE in grading, and providing a more precise catalogue of adverse effects in order to support the still subjective classification [Lacouture *et al.* 2011]. This subjectivity may be one major source for varying fractions of patients with a certain grading when comparing multiple studies.

Despite all discomfort, there is a positive aspect on skin rash as an adverse event: similar to the early description of EGFRI-induced skin reactions, a correlation between its severity and positive treatment outcome was reported even before FDA approval of Gefitinib [Herbst 2003]. Clinically similar observations have been made shortly after also for anti-EGFR antibodies, first in Cetuximab [Peréz-Soler & Saltz 2005; Lacouture 2006]. Anyhow, considering the two classes of EGFR-targeted classes of EGFRIs (mAbs and TKIs), differences with regard to incidence, severity and onset have been claimed [Melosky *et al.* 2009] so that this thesis, except as otherwise stated, relates to mAbs in general and Cetuximab in particular.

To date, the oncology community is deeply interested in this repetitively described connection between adverse effect intensity, drug activity and survival [Petrelli *et al.* 2013; Lupu *et al.* 2015]. For years now, skin toxicity is not only regarded as an undesired adverse effect, but also as pleasant phenotypical surrogate biomarker: the more severe the skin rash symptoms, the stronger anti-tumor effects, the better survival and the less progression [Petrelli *et al.* 2013]. Ultimately, these observations support decisions on optimized dose adaptions or switch of therapy [Leporini *et al.* 2013], while not requiring further invasive testing. Consequently, it is being used in the daily clinical practice – but to date, no full mechanistic link could be drawn on the molecular level. In other words: the regulative network and its molecular players remain hidden. While certain somatic alterations of EGFR itself, HLA-A and ABCG2 in the SR-negative patients have been described [Kozuki 2016], those are only explainatory for subpopulations of patients.

Apart from the academic question, the described implications show an important drawback for patients: neither drug efficacy nor skin rash severity are predictable, as this visual biomarker appears on treatment with EGFRIs only. Additionally, SR may be covered/silenced in the setting of preemptive skin care treatments [Leporini *et al.* 2013]. Therefore, patients have to suffer from likely discomfort and possibly severe skin reactions in order to be sure to undergo the appropriate treatment. There is

no clinical test at baseline yet, which could identify *a priori* the roughly 10% of patients who do not show any rash or only very mild forms and are likely to be EGFRI-resistant. For those patients, a treatment with EGFRIs is suggested as not beneficial [Petrelli *et al.* 2013], while valuable treatment time is lost.

However, several pieces of knowledge from dermatology, cancer research and immunology are already given, providing traces for shedding further light on the molecular genetic background, resulting in the stratification of patients into distinct reaction types.

1.4 Basic assumptions and rationals

In terms of biological tissues, both skin and colon are boundary layers to the environment and share some features. They separate the body and the outside, with digestion products being 'outside' by definition. Both show a constant directed growth in layers and dedicated aging from 'inside' to 'outside'. Accordingly, skin and colon are immunologically compentent and highly regulated. Diseases following a pathologic de-regulation of the respective mechanisms are well-known for both skin (e.g. psoriasis [Lowes *et al.* 2014]) and colon (inflammatory bowel diseases (IBD) [West & Jenkins 2015]), often being related to chronic inflammation.

Meanwhile, molecular mechanisms of inflammation are well known to be involved in carcinogenesis in general [Hanahan & Weinberg 2011; Pesic & Greten 2016]. For both skin [Sherwani *et al.* 2017] and colon [Kang & Martin 2017], knowledge on the interaction with the surrounding microbiota is accumulating: while the microbe composition influences the balance of immune system cycles, it gets increasingly recognized as important factor in the development of cancer as long-term progression of chronic inflammation. Even more, in several cancer settings a state of chronic inflammation has been reported to be triggered by the tumor itself or its microenvironment, supporting malignant growth including successful metastasizing [Smith & Kang 2013].

In turn, immunological reactions and carcinogenesis are closely connected to several cell adhesion mechanisms. These are highly regulated by pathways, including EGFR signaling. Like for the immune system (but also several other regulative molecular networks), only proper balancing of input signals prevents the system from pathologic states. Psoriasis may serve as a good example for pathologic misregulation of the skin, where immune reactions disturb proper cell adhesion. Interestingly, the EGFRI-induced skin rash symptoms are similar to such of autoimmune diseases like psoriasis [Peréz-Soler & Saltz 2005]. Conversely, for certain autoimmune diseases like systemic lupus erythematosus, drug induction has been described [Antonov *et al.* 2004; Vihinen *et al.* 2011]. Cetuximab being a strong

silencer of EGFR can be regarded as an extrinsic introducer of misregulation. Due to its systemic application, non-pathologic tissues in which EGFR plays a major role (like skin, hair and colon), are also affected.

Skin-located cancer cells are excluded as the cause of the rash as metastasizing from colon to cutaneous sides is very rare [Saladzinskas *et al.* 2010]. Although colon and skin share also a similar tumor biology, clinical phenotype, morphology, molecular patterns, and defects in cancer-related pathways [Berman 2004], there is no knowledge about direct connections between skin and colon tumor. However, the correlated clinical observations indicate Cetuximab to be responsible for both skin rash and antitumor effects on colon cancer lesions. Therefore, the different grade of skin rash between patients is assumed to be caused by minor naturally occurring germline variants somatically, not exclusively in the tumor sites (according to Kozuki 2016). Generally, this meets the concepts of pharmacogenetics/-genomics, considering individual genetic variance having an influence pathology as well as treatment reaction (genome-disease and genome-drug interactions), rising hope for individualized medicine [Ma & Lu 2011]. However, while in the light of 'omics data such knowledge is rapidly accumulating for tumor variations, especially in relatively well-studied entities like CRC [Bignucolo *et al.* 2017], systemic consideration of germline variations remains elusive. Interestingly, for general drug-induced skin adverse reactions (SCARs), genetic associations the major histocompatibility complex (MHC) genes have been reviewed very recently, although with the aim to support SCAR prevention [Gerogianni *et al.* 2018].

In a nutshell, the focus is set to naturally occurring germline variants in proteins and pathways which are expected to belong to one or more of three major functional groups: signaling, immune system and cell adhesion. The investigations aim to identify candidate variations potentially causing the skin rash phenomenon in Cetuximab-treated patients on the scale of molecular mechanisms.

1.5 Molecular networks and visualization

In the given setting, an unknown complex molecular mechanism regulates tissue reactions upon drug application, experienced as adverse effects. For a minority of patients, this mechanism is changed, resulting in an alternate behaviour. In terms of modeling, functional players (= proteins) and their relations (functional interactions) translate best into a network of nodes and edges. Within this regulated network, subsets of nodes and relations make up "pathways", mostly understood as signaling cascades utilizing a set of molecules, triggering certain effects upon certain signals. As regulation occurs via many cross-reactions between those pathways, the isolated view on defined pathways has to be taken as artificially simplified according to biology [Villaveces *et al.* 2015]. While

this may be applicable in tightly sketched scenarios on well- known pathways, incomplete knowledge on signaling cascades requires a more general strategy. Here, networks on curated protein-protein interaction (PPI) data offer a more complete view, although they do not take into account pathway knowledge [Villaveces *et al.* 2015]. For example, while two proteins A and B are interacting as well as B interacts with another protein C, there is no information contained whether A signals to C through B in any situation (= no transitivity). Such information has to be inferred from secondary resources like measurements or common knowledge.

Several approaches deal with quantitative data (mostly gene expression; e.g. [Lamb *et al.* 2006]), enabling mathematical modeling and the computation of dedicated scores (e.g. [Li *et al.* 2009b]). Furthermore, weighting of interactions may depend on e.g. assay ratings, probability estimations from empirical lab data and voting systems, as multiple references might be taken into account. In turn, various databases for interaction data[1] take into account multiple resources at a time (incl. biochemical assays, association studys and text mining on literature). Accordingly, a current trend in data analysis is the integration of various, case-specific high-throughput measurements, coining the term of "multi-omics" approaches [Huang *et al.* 2017].

Despite the arising complexity, those highly data-driven models appear attractive in terms of precision and exact mechanistic decription, and furthermore, they carry the chance to detect connections *de novo*. Anyhow, in the given use case, they are neither applicable nor necessary. This is on the one hand simply due to the given amount and type of data: exome data of 23 patients does not provide quantitative information. On the other hand, the aim is anyhow to support non-computer scientists in hypothesis generation. Here, field expert's background knowledge is intended to compensate the sparse information on the data level, designating this human-driven approach to act on level of high abstraction. In contrast to mathematical models, interfacing data and human experts is of central importance. Intuitively, the better data is presented, the easier it can be interpreted by connection with learned knowledge and personal experience (conclusions/inference). The concept of unconscious inference as central element in human visual perception has been introduced already in the 19th century and matters of making assumptions and conclusions from incomplete data, depending on prior knowledge [Helmholtz 1867]. Accordingly, proper visualization is key for hypothesis generation upon sparse data, which is anyhow expected to contain hidden patterns. For humans as visual-centric species [Nat. Methods Editorial 2005], the brain's perception capacities are exhausted soon in the case of serial information like full text, lists or tables [Duke *et al.* 2015]. In contrast, images can carry much

[1] For an extensive list, compare https://omictools.com/ppis-category

more information of high complexity [Few 2013]. Despite all intuitivity, the role of visual perception in particular has been formal subject of research, aiming for effective data visualization [Dastani 2002].

Apart from the applicability of graph algorithms in network analysis, networks instantly serve as easy and powerful visualizations. While the representation of entities and relations is intuitive, visualization complexity increases with the network's size, although various techniques like node filtering or layout algorithms to counteract this phenomenon [Villaveces *et al.* 2015]. Taken together, terms like "visual analytics" or "visual data mining" become common, independent of the particular subject of study (e.g. asthma [Bhavnani *et al.* 2014], gastric cancer [Pastrello *et al.* 2014] or Parkinson's disease [Porras *et al.* 2015]). Recently, in the face of the publically available masses of interaction data, awareness for usecase-specific manual curation rises, focussing on e.g. a disease or a particular gene of interest [Porras *et al.* 2015]. Aims may be the discovery of new biomarkers [Bhavnani *et al.* 2014], drug design [Sukumar & Krein 2012] or generally the extrapolation of relevant information from data [Pastrello *et al.* 2014]. In other words,: hypothesis generation. Here, interactivity is tremendously important. Recently, in a cancer genomics setting it could be proofed that there is a significant positive effect in enabling clinicians to interact with genomic reports, resulting in increased user satisfaction and improved data interpretation [Gray *et al.* 2018]. The authors clearly state to integrate interactive reports into electronic healthcare records, effectively facilitating the implementation of precision oncology. Analogously, from the perspective of future funding efforts for the use of big data in general health research, awareness of the importance of visualization has recently been demonstrated [Auffray *et al.* 2016]. Here, for progressing from time-consuming ranking lists and incomprehensible "hairball" networks to serious clinical decision support system, disease-specific maps have been explicitly suggested to be a considerable way.

1.6 Aim

The causal relation between adverse effects and anti-tumor drug efficacy in Cetuximab-treated CRC cases is still unknown, but of elevated interest in the context of clinical practice. Starting from a small set of selected patients, their somatic sequencing data should be mapped to a disease-specific network, depicting protein interactions. This map, being both graphically visualized and interactive, aims to support clinical field experts to develop hypotheses on candidate genes and mechanisms to investigate further. Beyond the biomedical interpretation of the overall results, the principle of the general “Molecular Systems Map” (MSM) and its usecase-specific application represents the core of this thesis.

2. Materials & Methods

The strategy for the overall work (Fig. 4) starts at two separate points: on the one hand, this is the patient-related exome sequencing data, which has to be converted into gene-centric variation information. This process is facilitated in Galaxy (comp. Section 2.3), GEMINI (comp. Section 2.4) and a range of custom shell commands and scripts (comp. Sections 2.5 and 2.6). On the other hand, the mechanistic context of the usecase from the literature has to be formalized into a network model, which is called Molecular Systems Map (MSM; comp. Subsection 2.7.1). This modeling as well as all subsequent steps are implemented in Cytoscape. Being the intercept point of the two initially disjoint branches, the mapping of the genes annotated in the variant calling to the MSM is the central step (comp. Subsection 2.7.2). The in such way populated MSM enables both visual inspection by a field expert and the application of graph algorithms, making up the basis for interactive hypothesis generation (comp. Section 4.3ff).

2.1 Patient overview

For the performed analyses, data of patients from two clinical studies where used. Overall, 7 and 16 patients were selected from the CIOX (or Fire2; [Moosmann *et al.* 2011]) and FIRE3 [Heinemann *et al.* 2014] study, respectively. Apart from clinical and epidemiologic parameters, exome datasets (raw paired end short reads in FASTQ format) from 23 patients were provided and used for the initial data analysis and generation of the model (MSM; comp. Section 2.7).

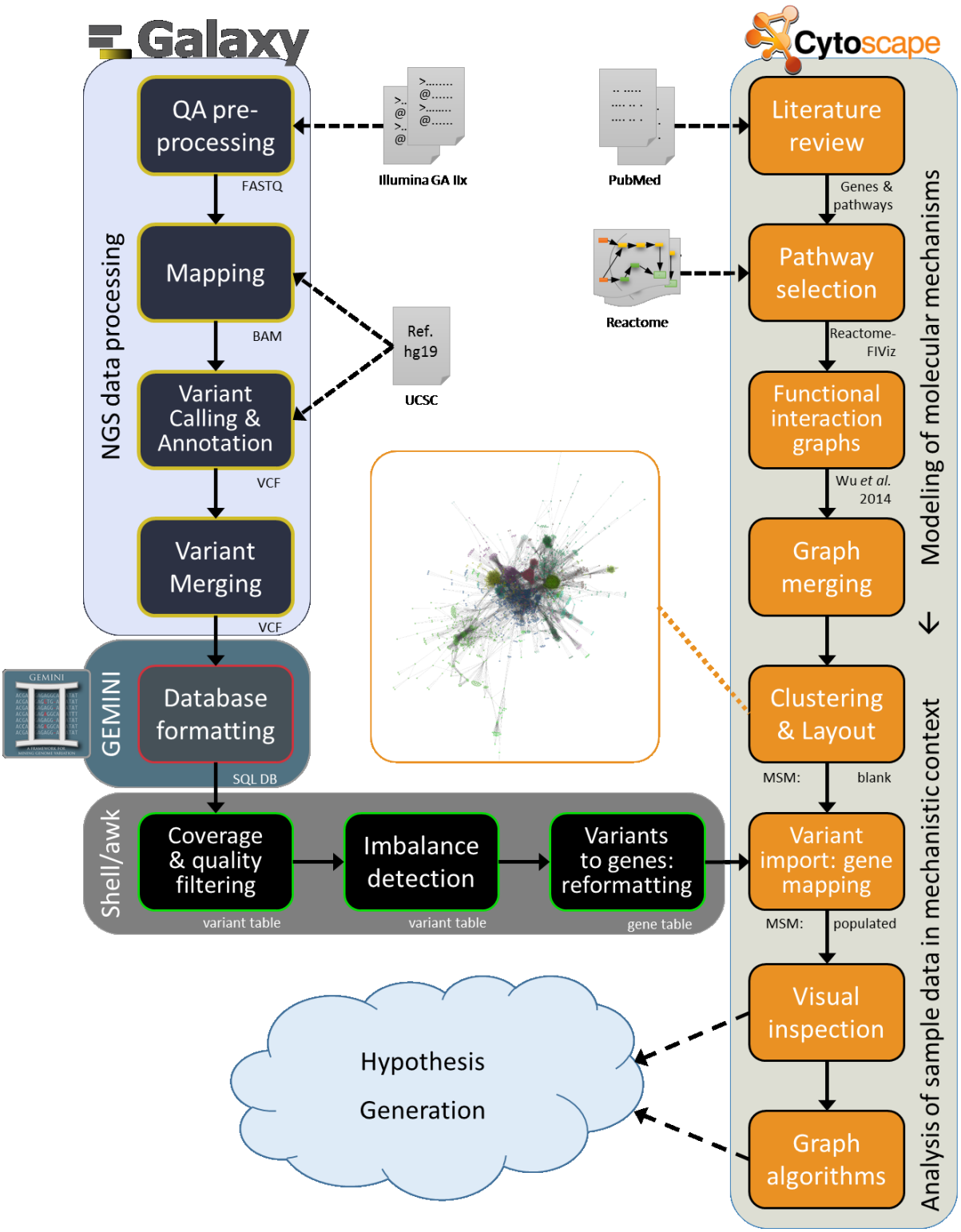


Fig. 4 - Depiction of the thesis overall workflow. Two branches, processing the NGS data to gene-centric variants of interest (left) and modeling the mechanistic context of the usecase according to public knowledge (but also discussion with field experts; right), are joined by mapping variant-affected genes to the newly developed central data structure, the Molecular Systems Map (MSM; middle). Major components of the left branch are Galaxy for NGS short read processing, GEMINI for genomic variant storage and querying, as well as a set of shell commands and custom scripts, primary in awk, but also R. A reference genome is necessary. The right branch and all subsequent steps, especially the user-driven ones, are hosted by Cytoscape, ultimately supporting interactive hypothesis generation, taking into account both actual NGS datasets and functional relations of mechanisms of interest. Basis for the modeling phase is pathway information from Reactome.

Tab. 1 - Overview on study patients. Exome data had been generated from the following patients (case report forms/CRFs for identification) out of two mCRC studies [Moosmann *et al.* 2011; Heinemann *et al.* 2014], where Cetuximab has been applied. Skin rash (SR) grades following NCI CTCAE v3.0 [Trotti *et al.* 2003]. Lab ID and CRF are equal for Fire3 patients, while individual labels exist for CIOX. SR = 0/1 indicates for 'negative', SR = 2/3 for 'positive' group. PFS = progression-free survival, OS = overall survival, ORR = overall response rate.

Study	CRF	Lab ID	SR grade	Age	Sex	KRAS WT	PFS [months]	OS [months]	ORR
CIOX	22	111	2	66	m	no	6.3	30.2	0
CIOX	67	155	3	56	m	yes	7.0	36.0	1
CIOX	110	014	3	68	m	yes	8.6	19.4	1
CIOX	114	072	0	66	f	yes	4.6	17.0	0
CIOX	129	036	0	71	m	no	1.9	4.9	0
CIOX	163	125	0	51	f	yes	2.8	11.0	0
CIOX	167	137	3	75	m	no	5.5	17.9	1
FIRE-3	20	020	1	71	m	no	2.9	13.6	0
FIRE-3	90	090	3	57	f	unknown	15.3	63.0	1
FIRE-3	213	213	3	42	m	yes	61.1	63.5	0
FIRE-3	281	281	1	68	f	no	1.4	19.2	0
FIRE-3	344	344	3	73	m	no	63.1	63.1	1
FIRE-3	375	375	1	67	f	no	1.5	5.0	0
FIRE-3	406	406	1	65	f	yes	28.4	55.6	0
FIRE-3	428	428	1	73	m	yes	3.7	3.7	0
FIRE-3	566	566	3	64	m	yes	14.5	14.7	1
FIRE-3	586	586	1	59	f	yes	1.9	12.9	0
FIRE-3	598	598	3	56	m	yes	28.5	28.5	1
FIRE-3	624	624	3	73	m	yes	3.9	14.2	0
FIRE-3	638	638	3	61	f	yes	4.9	30.9	0
FIRE-3	708	708	3	43	m	yes	12.5	27.2	1
FIRE-3	750	750	1	67	f	no	1.5	4.3	0
FIRE-3	796	796	1	73	f	yes	1.4	16.0	0

2.2 Exome data

Generation of exome data has been performed ahead of the presented work, but dedicated to uncovering the reasons for the skin rash phenomenon. Between April and December 2012, the data were produced at the LMU Gene Center, applying an Illumina Genome Analyzer IIx, which received a capacity upgrade in this period. Therefore, multiplexing was introduced for later samples. In consequence, the single datasets slightly differ in terms of coverage and quality (comp. Tab. 2). For exome targeting, the Agilent SureSelect Human all exon 50Mb kit was used. Raw data sets are listed in Tab. 2, a basic depiction of sequence count vs. read lengths in Fig. 5.

Tab. 2 - Raw datasets generated at the LMU Gene Center. Data were generated between April and December 2012 as paired-end reads on overall four chips. In between, the device was upgraded in terms of sequencing capacity, enabling multiplexing. Therefore, on later chips samples were distributed across multiple lanes (three or five, respectively), the resulting FASTQ files simply concatenated into one each. Descriptive statistics apply for those combined files. Theoretical coverage has been calculated assuming an equal distribution of sequenced bases (read count * read length) on 50 Mb target sequence.

Pat. ID Lab	Sequencing date	Chip ID	No. of lanes	forward/ reverse [F/R]	total sequence count	read length [bp]	GC content [%]	theo. coverage
111	2012-04-10	HWUSI-EAS632R_00054	1	F	49,357,115	85	46	84x
				R	49,357,115	85	46	84x
155	2012-04-10	HWUSI-EAS632R_00054	1	F	50,868,088	85	47	86x
				R	50,868,088	85	47	86x
014	2012-04-10	HWUSI-EAS632R_00054	1	F	47,536,362	85	51	81x
				R	47,536,362	85	51	81x
072	2012-05-31	HWUSI-EAS1783R_00005	1	F	65,239,385	76	50	99x
				R	65,239,385	84	50	110x
036	2012-04-10	HWUSI-EAS632R_00054	1	F	50,578,517	85	48	86x
				R	50,578,517	85	48	86x
125	2012-05-31	HWUSI-EAS1783R_00005	1	F	49,236,313	76	48	75x
				R	49,236,313	84	48	83x
137	2012-05-31	HWUSI-EAS1783R_00005	1	F	57,888,647	76	43	88x
				R	57,888,647	84	43	97x
020	2012-12-13	HWUSI-EAS1783R_00020	5	F	31,002,461	76	49	47x
				R	31,002,461	84	49	52x
090	2012-11-29	HWUSI-EAS1783R_00018	3	F	21,603,701	76	46	33x
				R	21,603,701	84	46	36x
213	2012-12-13	HWUSI-EAS1783R_00020	5	F	24,779,449	76	46	38x
				R	24,779,449	84	46	42x
281	2012-12-13	HWUSI-EAS1783R_00020	5	F	29,490,809	76	46	45x
				R	29,490,809	84	46	50x
344	2012-11-29	HWUSI-EAS1783R_00018	3	F	21,585,819	76	46	33x
				R	21,585,819	84	46	36x
375	2012-12-13	HWUSI-EAS1783R_00020	5	F	27,631,902	76	48	42x
				R	27,631,902	84	48	46x
406	2012-12-13	HWUSI-EAS1783R_00020	5	F	33,694,938	76	50	51x
				R	33,694,938	84	50	57x
428	2012-12-13	HWUSI-EAS1783R_00020	5	F	30,022,010	76	48	46x
				R	30,022,010	84	48	50x
566	2012-11-29	HWUSI-EAS1783R_00018	3	F	24,328,044	76	46	37x
				R	24,328,044	84	46	41x
586	2012-11-29	HWUSI-EAS1783R_00018	3	F	21,175,515	76	45	32x
				R	21,175,515	84	45	36x
598	2012-11-29	HWUSI-EAS1783R_00018	3	F	22,767,044	76	46	35x
				R	22,767,044	84	45	38x
624	2012-12-13	HWUSI-EAS1783R_00020	5	F	22,403,986	76	46	34x
				R	22,403,986	84	46	38x
638	2012-11-29	HWUSI-EAS1783R_00018	3	F	24,637,690	76	45	37x
				R	24,637,690	84	45	41x

708	2012-12-13	HWUSI-EAS1783R_00020	5	F	23,165,578	76	46	35x
				R	23,165,578	84	46	39x
750	2012-12-13	HWUSI-EAS1783R_00020	5	F	25,770,359	76	46	39x
				R	25,770,359	84	46	43x
796	2012-12-13	HWUSI-EAS1783R_00020	5	F	23,058,518	76	46	35x
				R	23,058,518	84	46	39x

As shown, both read count and GC content are equal when comparing forward and reverse read sets of a single patient. Read count and lengths were sufficient, and FastQC warnings could be overcome by the first, QA-related steps of the processing pipeline (comp. Section 2.3), maily due to simply using the quality-based trimmer. Taken together, data were considered to be valid for further use.

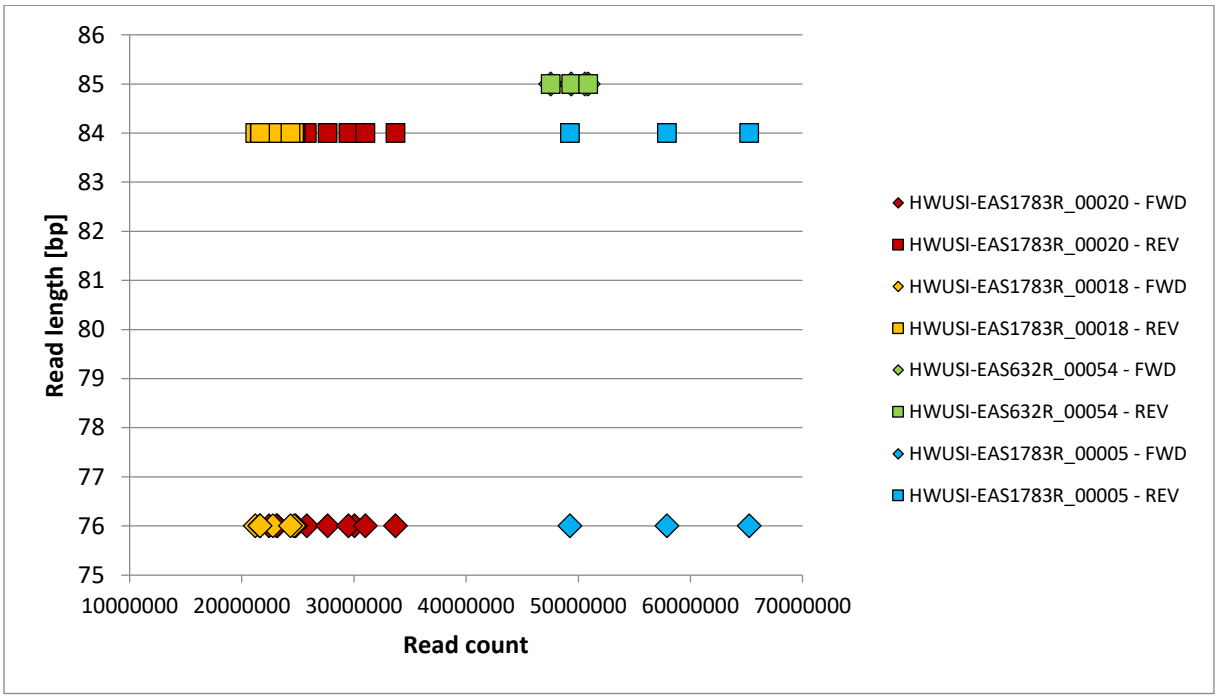


Fig. 5 - Read count vs. read length plot for all forward and reverse datasets, separated by chip. Except for one chip (green), where they are equal, read lengths for forward reads are 8 bp shorter than for reverse ones, presumably by lab design. For the older chips, a single lane delivers roughly double the data compared to the newer, multiplexed ones, although they use three or five (shared) lanes per sample, respectively. Meanwhile, the lengths are equal or higher for the non-multiplexed chips.

2.3 From raw reads to variant calling

For input datasets, where reads from exome sequencing of a biological sample were distributed across multiple files (due to multiplexing via DNA barcoding), collections of forward and reverse raw reads, respectively, were merged first. The simple merges are used as input datasets in the variant calling pipeline.

In a multistep process, the collected paired-end raw reads were translated into variants. This pipeline, set up in the local Galaxy [Afgan *et al.* 2016] instance of the Medical Faculty of the Ludwig Maximilians University Munich ("NGS-FabLab" [Schaaf *et al.* 2014]), principally consists of a quality assurance and control (QA/QC), a short read mapping to the human reference genome hg19, a variant calling using three algorithms and a final genomic annotation of the variants. For raw and trimmed reads (forward and reverse each), initial bwa mapping and final re-alignment of reads, FastQC was applied in order to receive summary reports on quality measures (comp. Fig. 6). For the finally re-aligned reads, Picard's 'Collect Alignment Summary Metrics' module[2] was applied additionally. Variant calls were performed using VarScan2 (SNP and InDel algorithm each [Koboldt *et al.* 2012]) and GATK's HaplotypeCaller[3], merged afterwards and finally annotated with SNPeff [Cingolani *et al.* 2012] (comp. Fig. 7). Most of the applied modules are part of either the samtools [Li *et al.* 2009a], Picard[4] or GATK v2.8[5] toolboxes.

By application of the pipeline, for each patient one Variant Call Format (VCFv4.1[6]) file was generated, containing three technical samples each (one sample column per variant caller result).

For an exact listing of command lines executed by Galaxy, refer to Boxes Box A2 to Box A4 in the Appendix. For those commands applied for generating the QC plots shown in Section 3.2, refer to Boxes Box A5 to Box A8.

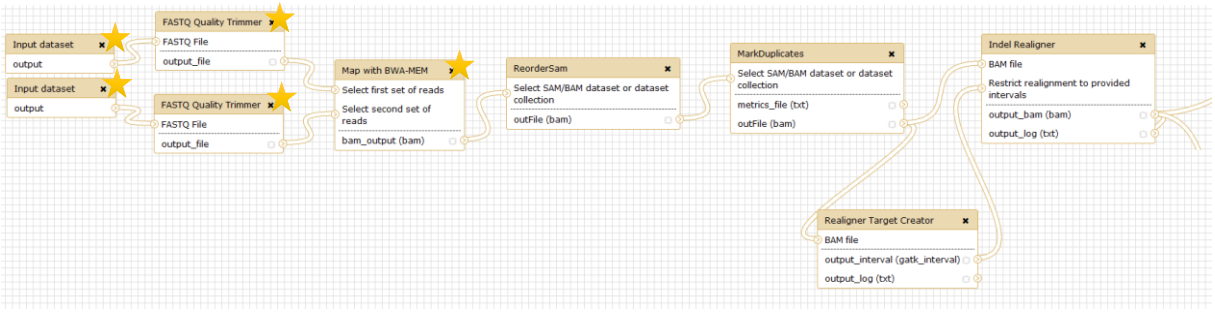


Fig. 6 - Galaxy workflow view (1/2) on exome variant calling pipeline applied to patient samples (continued in Fig. 7). Quality-based trimming of ends was followed by mapping to the hg19 reference genome, reordering, duplicate marking and indel re-alignment. Outputs of modules with yellow stars underwent quality control using FastQC.

[2] https://broadinstitute.github.io/picard/command-line-overview.html#CollectAlignmentSummaryMetrics
[3] https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_haplotypecaller_HaplotypeCaller.php
[4] http://broadinstitute.github.io/picard/
[5] https://github.com/broadgsa/gatk-protected/tree/2.8
[6] https://samtools.github.io/hts-specs/VCFv4.1.pdf

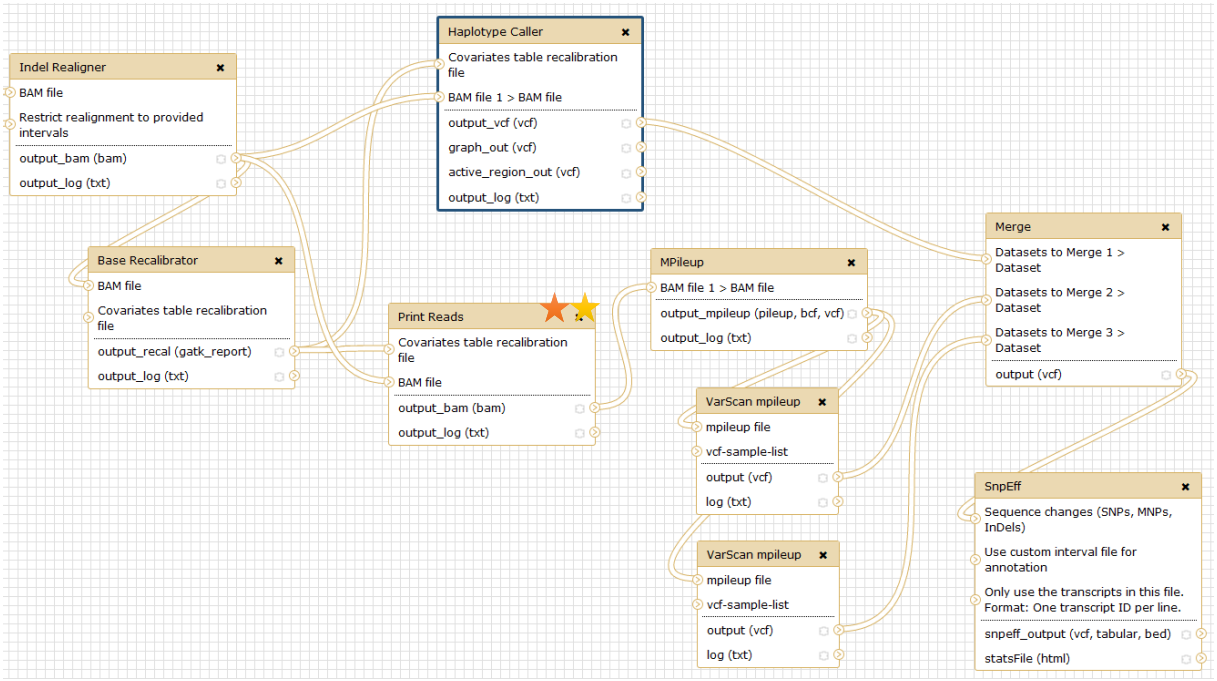


Fig. 7 - Galaxy workflow view (2/2) on exome variant calling pipeline applied to patient samples (continuation of Fig. 6)**.** The 'Indel Realigner" box is the same like in Fig. 6, linking both figures. After re-alignment and base recalibration, in one branch GATK's Haplotype Caller, in the other branch both VarScan's SNP and InDel implementations are applied. Results of all three variant caller runs are merged and annotated with SNPeff. Outputs of modules with yellow stars underwent quality control using FastQC, Picard's 'Collect Alignment Summary Metrics' module was applied to the last version of the BAM-formatted mapping data (orange star).

2.4 Post-processing of VCF files

For the centralized and queriable storage of called variants, GEMINI v0.18.2 [Paila *et al.* 2013] was chosen. According to the official manual[7], VCF files were separately decomposed (Box 1) and normalized (Box 2) with 'vt' [Tan *et al.* 2015], followed by sorting by genomic position (VCFTools v0.1 [Danecek *et al.* 2011]) and indexing (tabix v0.2.5[8]; from samtools; Box 3) in order to merge all files (Box 4). After merging, decomposition and normalization was repeated.

Box 1 - Looped, 'vt'-based decomposition of patient VCF files.

```
$ cd exome_vcf/
$ mkdir -p preprocessing/decomposed
$ for f in `ls -1p | grep -v "/"`; do
    vt decompose -s $f -o preprocessing/decomposed/$f
  done
```

[7] https://gemini.readthedocs.io/en/latest/#new-gemini-workflow
[8] http://www.htslib.org/doc/tabix.html

Box 2 - Looped, 'vt'-based normalization of decomposed VCF files.

```
$ mkdir -p preprocessing/normalized
$ for f in `ls -1 preprocessing/decomposed/`; do
      vt normalize -r [path]/hg19.fa $f \
      > preprocessing/normalized/`basename $f`
  done
```

Box 3 - Looped sorting, zipping and indexing of decomposed and normalized patient VCF files.

```
$ PERL5LIB=/home/galaxy/galaxy/tool_dependencies/vcftools/0.1.11/\
   devteam/package_vcftools_0_1_11/710efaae2ff8/vcftools/lib/\
   perl5/site_perl:$PERL5LIB
$ export PERL5LIB
$ wd=preprocessing/sorted_zipped_indexed
$ mkdir -p $wd
$ cd $wd
$ for f in `ls -1 ../normalized/*.vcf`; do
    outname=`basename $f .vcf`.sorted.vcf
    echo $outname
    vcf-sort $f > $outname
    gzip $outname
    tabix -p vcf ${outname}.gz
  done
```

Box 4 - Merging of all pre-processed patient-related VCF files, followed by additional decomposition and normalization.

```
$ dir=../merged
$ mkdir -p $dir
$ vcf-merge *.gz > $dir/all_merged.vcf
$ cd $dir
$ vt decompose -s all_merged.vcf \
               -o `basename all_merged .vcf`.decomposed.vcf
$ vt normalize -r /EasyRaid/share/genomes/links/hg19.fa \
      `basename all_merged .vcf`.decomposed.vcf \
      > `basename all_merged.decomposed.vcf .vcf`.normalized.vcf
$ cd ../../
```

The GEMINI database was loaded (initialized) in two ways: on the one hand once with every pre-processed patient VCF (resulting in 23 separate DBs), on the other hand with the merged VCF file (resulting in one DB containing data of all patient variants). While the 23 separate patient DBs were used for QC purposes (see Section 2.5) only, all other analyses were performed using the complete merged database holding all patient samples.

The background is that the measure of comparing the VCF 'QUAL' column with the sample-specific quality value ('gt_quals' in GEMINI corresponding to VCF 'GQ' tag) on a per-sample basis is impossible after merging, as the "QUAL" value is averaged across all included samples. Therefore, comparing QUAL to sample-specific values requires separate DBs.

Box 5 - GEMINI loading and annotation of sample information

```
$ export PYTHONPATH=[PATH_TO_ANACONDA]/lib/python2.7/:$PYTHONPATH
$ mkdir gemini_DBs
$ # create sample-specific DBs
$ for f in `ls -1 preprocessing/normalized/*.vcf`; do
      echo "PROCESSING FILE "$f"...";
      gemini load --cores 12 -v $f -t snpEff \
          gemini_DBs/`basename $f.db`
      gemini amend --sample all_pats_training.ped \
          gemini_DBs/`basename $f.db`
  done
$ # create complete DB from merged samples
$ gemini load --cores 12
  -v preprocessing/merged/all_merged.decomposed.normalized.vcf
  -t snpEff gemini_DBs/all_merged.decomposed.normalized.db
$ # annotate sample information via PED file
$ gemini amend --sample \
      all_pats_training_more_info_varcaller_triple.ped \
      gemini_DBs/all_merged.decomposed.normalized.db
```

2.5 Quality control of called variants by visual inspection

In order to identify failed samples, abnormal analyses and other outliers due to biological or technical artifacts, a series of XY scatterplots was generated from the GEMINI databases.

First, supportive attributes were chosen, namely the chromosomal ‘start’ position (‘POS’ in VCF) and ‘quals’ (VCF: ’QUAL’), both being variant-specific, as well as sample-specific ‘gt_depths’ and ‘gt_quals’ (VCF ‘DP’ and ‘GQ’ field, respectively). As the output of one variant caller results in trailing sample column each within the patient-specific VCF file, “sample” means “variant caller” in the given setting.

Second, for all cross-pairings collections of images were generated by applying a self-developed R script, which queries the database and processes the outputs. The resulting plot collections were pasted together making up panels with chromosomes (23; X axis) vs. patients (23; Y axis). For sample-specific attributes, marks were color-coded according to the samples (with “sample” again meaning “variant caller”). Parts of the plot collections were additionally generated with logarithmic scales.

According to the plots and referring to the meaning of the PHRED score, minimal thresholds for ‘gt_depth’ and ‘gt_quals’ were chosen to be 5 and 13, respectively. That is, a variant site has to be covered by five or more short reads, with a variant being detected at a 95% accuracy or higher (variant caller internal statistic). These relatively permissive values take into account the overall coverage and quality of the initial NGS data, which is in comparison to nowadays datasets relatively low.

2.6 Pre-candidate selection: imbalance measure and gene mapping

In order to overcome statistical limitations considering the small number of patients, the simple approach of detecting a variant, which is distributed between both patient groups in an imbalanced manner, was chosen.

2.6.1 Statistical assumptions on imbalanced variants

Statistically expressed, the basic assumption is that a variant's appearance in patients is assumed to occur in a uniform and independent pattern. By regarding this setting as a series of trials (patients of a group), one can describe a general Bernoulli process, knowing the states 'success' (variant present) and 'failure' (variant absent) – independent from the exact pattern/sequence across individual patients. Consequently, for a *k* describing the number of occurrences of a variant within a group of patients, a binomial distribution can be modelled, depending on *n* (the number of patients) and *m* (the allele frequency). Taken together, the probability *B* can be formalized as:

$$B(\,k \mid m,n\,) = \binom{n}{k} m^k (1-m)^{n-k} \quad \text{for} \quad k = 0,1,\dots,n$$

For the imbalance overall, the probability *P*(*imb*) is conditionally cummulative, summing up the products of probabilies for every possible number of variant-affected patients of the one condition with every other possible variant-affected patient count in the other group. Here, the imbalance criterion *t* restricts the set of accepted cases by expressing a minimal difference in success (presence) counts between the groups:

$$P(imb,t) = \sum_{i=0}^{n}\sum_{j=0}^{p} x_{i,j} \qquad \text{where}$$

$$x_{i,j} = \begin{cases} B(i \mid b,n)\cdot B(j \mid b,p) & |i-j| \geq t \\ 0 & \text{otherwise} \end{cases}$$

with *n* and *p* being the number of patients in the SR-negative and SR-positive group, respectively.

Following this probability model, for detected variants their imbalance measures, Minor Allele Frequencies (MAF; *a*) and genotypes were used to plot distributions (comp. Fig. 16 and Fig. 17), with their maxima depicting the minimal probabilities for coincident appearance of an imbalanced variant (i.e., considered to be insignificant for the SR use case).

$$b = \begin{cases} a^2 & \text{homozygous variant} \\ 2a\,(1-a)\, + a^2 & \text{otherwise} \end{cases}$$

While the two alleles *A'* and *A''* of a gene in diploid organisms can be assumed to be independent, the MAF *a* is to be squared when considering homozygous variants, making up the first case for *b*. When including also heterozygous variants (which in turn would require to be of dominant trait when considering them to be phenotypically effective), the probabilities for both the case that *A'* is altered while *A''* is not and *vice versa* have to be considered, making up the second case for *b*.

2.6.2 Imbalance measure on called variants and proceeding to genes

For retreiving the variants and their distribution in the patient groups, SQL statements were formulated in order to query the GEMINI database. The statements join information from both the 'variants' and the 'gene_detailed' table, achieving to include information about affected transcripts and both IDs from HGNC and Entrez for later gene symbol normalization. Considering optional restrictions to homozygous variants and the filtering of low impact variants (like intronic), four queries were executed (comp. Box 6 to Box 9).

Box 6 - Detection of imbalanced variants (Set 1): heterozygous included, all variants independent of impact strength.

```
$ gemini query --header -q \
  "SELECT v.variant_id, v.chrom, v.start, v.ref, v.alt, v.rs_ids, \
   v.aaf_1kg_all, v.impact_so, v.impact_severity, v.gene, \
   g.transcript, g.is_hgnc, g.hgnc_id, g.entrez_id, \
   (gt_quals).(phenotype==0), (gt_quals).(phenotype==1), \
   (gt_depths).(phenotype==0), (gt_depths).(phenotype==1), \
   (gt_types).(phenotype==0), (gt_types).(phenotype==1), \
   (gts).(phenotype==0), (gts).(phenotype==1) \
   FROM variants v, gene_detailed g \
   WHERE v.chrom = g.chrom AND v.gene = g.gene \
   ORDER BY v.gene ASC, v.variant_id ASC" \
   ../gemini_dbs/all_merged.decomposed.normalized.vcf.db | \
   awk -v offset=14 -v minq=13 -v minc=5 \
     -f gemini_imbalance_detector.awk \
     >> imbalance_7_HET_final.ssv
```

Box 7 - Detection of imbalanced variants (Set 2): heterozygous included, variants annotated as 'LOW' impact by SNPeff are ignored.

```
$ gemini query --header -q \
  "SELECT v.variant_id, v.chrom, v.start, v.ref, v.alt, v.rs_ids, \
   v.aaf_1kg_all, v.impact_so, v.impact_severity, v.gene, \
   g.transcript, g.is_hgnc, g.hgnc_id, g.entrez_id, \
   (gt_quals).(phenotype==0), (gt_quals).(phenotype==1), \
   (gt_depths).(phenotype==0), (gt_depths).(phenotype==1), \
   (gt_types).(phenotype==0), (gt_types).(phenotype==1), \
   (gts).(phenotype==0), (gts).(phenotype==1) \
   FROM variants v, gene_detailed g \
   WHERE v.chrom = g.chrom AND v.gene = g.gene \
   AND impact_severity != 'LOW' \
   ORDER BY v.gene ASC, v.variant_id ASC" \
   ../gemini_dbs/all_merged.decomposed.normalized.vcf.db | \
   awk -v offset=14 -v minq=13 -v minc=5 \
     -f gemini_imbalance_detector.awk \
     >> imbalance_7_HET_noLOW_final.ssv
```

Box 8 - Detection of imbalanced variants (Set 3): homozygous only, all variants independent of impact strength.

```
$ gemini query --header -q \
  "SELECT v.variant_id, v.chrom, v.start, v.ref, v.alt, v.rs_ids, \
   v.aaf_1kg_all, v.impact_so, v.impact_severity, v.gene, \
   g.transcript, g.is_hgnc, g.hgnc_id, g.entrez_id, \
   (gt_quals).(phenotype==0), (gt_quals).(phenotype==1), \
   (gt_depths).(phenotype==0), (gt_depths).(phenotype==1), \
   (gt_types).(phenotype==0), (gt_types).(phenotype==1), \
   (gts).(phenotype==0), (gts).(phenotype==1) \
   FROM variants v, gene_detailed g \
   WHERE v.chrom = g.chrom AND v.gene = g.gene \
   ORDER BY v.gene ASC, v.variant_id ASC" \
   ../gemini_dbs/all_merged.decomposed.normalized.vcf.db | \
   awk -v offset=14 -v rfilter=1 -v minq=13 -v minc=5 \
     -f gemini_imbalance_detector.awk \
     >> imbalance_7_HOM_final.ssv
```

Box 9 - Detection of imbalanced variants (Set 4): homozygous only, variants annotated as 'LOW' impact by SNPeff are ignored.

```
$ gemini query --header -q \
  "SELECT v.variant_id, v.chrom, v.start, v.ref, v.alt, v.rs_ids, \
   v.aaf_1kg_all, v.impact_so, v.impact_severity, v.gene, \
   g.transcript, g.is_hgnc, g.hgnc_id, g.entrez_id, \
   (gt_quals).(phenotype==0), (gt_quals).(phenotype==1), \
   (gt_depths).(phenotype==0), (gt_depths).(phenotype==1), \
   (gt_types).(phenotype==0), (gt_types).(phenotype==1), \
   (gts).(phenotype==0), (gts).(phenotype==1) \
   FROM variants v, gene_detailed g \
   WHERE v.chrom = g.chrom AND v.gene = g.gene \
   AND impact_severity != 'LOW' \
   ORDER BY v.gene ASC, v.variant_id ASC" \
   ../gemini_dbs/all_merged.decomposed.normalized.vcf.db | \
   awk -v offset=14 -v rfilter=1 -v minq=13 -v minc=5 \
     -f gemini_imbalance_detector.awk \
     >> imbalance_7_HOM_noLOW_final.ssv
```

The ‘gemini_imbalance_detector’ has been written in `awk` [Aho *et al.* 1988], providing an easy, line-wise processing of the query outputs at a maximum performance. For this tool, the minimal thresholds for coverage/depths (‘minc’) and quality (‘minq’) are pre-configured, although configurable.

The detector takes three consecutive columns per patient and acknowledges a variant if at least one of the three variant callers passes the depth and quality filters. Outputs are space-separated files (.ssv). Further documentation is provided within the script itself.

As indexes and databases of bwa, the variant callers and SNPeff may not be up to date compared to subsequent tools (esp. ReactomeFI; comp. 2.7.1), gene symbols were updated by comparison to a recent HGNC [White *et al.* 1997] excerpt. By using the HGNC ID, preferred/approved names were appended as additional column to the imbalance detector outputs, performed by the ‘HGNC_approver’ script (comp. Box 10). From here on, all files are tab-separated (.tsv).

In a subsequent two-step process, the GEMINI database outputs were further processed by merging lines according to (1) transcripts and (2) genes. The ‘line_merger’ script takes columns (by name) to check for equality in consecutive lines. For those line sets, the contents of columns configured to merged are pasted together, using a configurable separator symbol.

For collecting transcripts (and e.g the respective predicted effects), lines with both the same variant ID and gene were merged, resulting in comma-separated lists for transcript-specific attributes. Afterwards, lines referring to the same gene were merged according to the variant information, as one gene may be affected by multiple variants according to SNPeff. For later steps, a per-gene treatment of data is necessary.

In order to have explicit information of the number of merged lines, the 'line_merger' script (Box 10) was configured to append the columns "affectedTranscriptCount" and "variantsCountForGene", respectively.

Box 10 - Processing of the imbalanced GEMINI outputs (with variable '$file' standing for the four outputs of Box 6 to Box 9). Normalization of gene symbols according to HGNC, merging of variant lines affecting multiple transcripts as well as merging variants affecting the same gene (per-gene representation).

```
# normalize to HGNC-approved name (appends column "approved_name")
$ awk -F $'\t' -f HGNC_approver.awk HGNC_all.tsv $file \
   >> $normed_file;

# merge lines by variant: each line contains a variant and
#  all transcripts being affected by it (attributes of the
#  latter in a comma-separated list per 'merge' field)
$ awk -F $'\t' \
      -v merge="transcript,impact_so,impact_severity" \
      -v check="variant_id,gene" \
      -v delim=";" \
      -v cntcol="affectedTranscriptCount" \
      -f line_merger.awk \
      $normed_file >> $merged_file1

# merge lines by gene: one line carries a gene and all variants
#  affecting it (attributes of the latter in a pipe-separated
#  list per 'merge' field))
# Define columns to check for equality first
$ check="chrom,gene,approved_symbol,is_hgnc,hgnc_id,entrez_id";

# Afterwards, construct a comma-separated list of all other
#  columns to merge
$ merge=`head -n1 $file | awk -v check=$check \
          'BEGIN{ split( check, ccols, "," );
                  for( c in ccols ){store[ccols[c]] = "T"};
                  sep=""
          }
          { for( i=1; i<=NF; i++){
              if(length(store[$i]) < 1){
                printf("%s%s", sep, $i);
                sep="," }
              }
          }
        END{ print "\n" }'`

# Finally, merge by using the line_merger script again, using
#  pipe ('|') as separator
$ awk -F $'\t' \
      -v merge=$merge \
      -v check=$check \
      -v delim="|" \
      -v cntcol="variantsCountForGene" \
      -f line_merger.awk \
      $merged_file1 >> $merged_file2;
```

2.7 Molecular Systems Map (MSM)

In a first step, the map is generated by collecting pathway (interaction) data, depending on the use case and joining the items to a network. In a second step, the MSM gets 'populated', i.e. those genes are identified, which are hit by at least one variant ('affected genes').

2.7.1 Generation of a blank MSM

As platform for the integration of pathway data and visualization, Cytoscape v3.4.0[9] [Shannon *et al.* 2003] was chosen. Additionally, the plugins ReactomeFIViz (v5.1.0.beta, [Wu *et al.* 2014]) and AllegroLayout v2.2.2[10] were installed via the integrated app manager. For shortest path detection, Pesca (v3.0.8[11]; [Scardoni *et al.* 2015]) was used.

From a 'Reactome Pathways' [Croft *et al.* 2014] session, diagrams were selected corresponding to the defined major topics "signaling", "immune system" and "cell adhesion", plus "cellular stress". Within the hierarchically organized tree structure of Reactome's interaction event collection, the outmost diagrams (leafs) of interest were selected and converted into a FI networks, each containing functional interaction data from Reactome and further sources [Wu *et al.* 2010; Wu *et al.* 2014] like KEGG. Those networks (pathways) were merged iteratively 'bottom-up', making up subsets, sets and finally supersets of interest (further details and description see Tab. 3). In order to track the origin or membership of nodes and edges, for each of the initial Reactome-derived networks, as well as any derived set, boolean columns were introduced and filled up with 'true' values. Therefore, nodes and/or edges can be selected in merged networks (incl. the MSM itself) according to annotated pathways.

For particular networks, nodes of excessively high degree were removed, if the expected biological impact for the given analysis was low. This affects in the very first line the multiple immuneglobulin chains and their respective gene loci. Those items are close to fully connected with each other and few further nodes, providing only generic information. Additionally, these loci are anyhow somatically altered in immune cells. Technically, the number of the adjacent edges to these nodes make up the vast majority of edges within the graphs and slow down Cytoscape to an unserviceable level. For a full list of these removed nodes refer to Tab. 4.

In a final step, the resulting three supersets together with the "cellular responses to stress" set were merged to the blank (means: carrying no sample-specific genetic data) MSM. The complete network

[9] http://www.cytoscape.org/release_notes_3_4_0.html
[10] became recently unavailable, compare https://groups.google.com/forum/#!topic/cytoscape-helpdesk/122pCTu_0AI
[11] http://apps.cytoscape.org/apps/pesca30

was re-ordered with AllegroLayout, using the Fruchterman-Reingold algorithm [Fruchterman & Reingold 1991] and default parameters (2,000 iterations, independent component processing, scale and gravity at 100% each). Finally, a functional clustering was calculated using the ReactomeFI plugin's internal feature (spectral partition based network clustering; [Newman 2006]), resulting in functional groups ("modules"). The session was saved as 'molecular_map_with_history.cys' plus a copy in which for performance reasons all non-MSM networks were removed ('molecular_map_only.cys').

Tab. 3 - Selected pathways according to ReactomeFI. Entries providing no diagram in Cytoscape's ReactomeFI app are printed in grey. Sets in italics show more subordinate pathways in Reatome, but not all have been extracted for the MSM. Supersets do not have correspondently named diagram in ReactomeFI. Classifications of mechanisms as sets, subsets etc. corresponding to relative depths within the hierachy. ReactomeFI diagrams which could not be further resolved to any subdiagrams were labeled as "pathway". Consequently, the subset level may be left empty.

Superset	Set	Subset	Pathway
Signal-transduction	Signaling by EGFR		
	Signaling by FGFR		Signaling by FGFR1 Signaling by FGFR2 Signaling by FGFR3 Signaling by FGFR4
	Signaling by VEGF		
	Signaling by ERBB2		
	Signaling by ERBB4		
	PIP3 activates AKT signaling		
	MAPK family signaling cascades	MAPK1/MAPK3 signaling	RAF/MAP kinase cascade RAF-independent MAPK1/3 activation
		MAPK6/MAPK4 signaling	
	Signaling by Rho GTPases		
	Signaling by TGF-beta Receptor Complex		
	Signaling by Wnt		WNT ligand biogenesis and trafficking Degradation of beta-catenin by the destruction complex TCF dependent signaling in response to WNT Beta-catenin independent WNT signaling
Immune System	*Adaptive Immune System*		TCR signaling Costimulation by the CD28 family Signaling by the B Cell Receptor (BCR) Class I MHC mediated antigen processing & presentation MHC class II antigen presentation Immunoregulatory interactions between lymphoid and non-lymphoid cells
	Innate Immune System		Toll-like Receptors Cascades Complement cascade Nucleotide-binding domain, leucine rich repeat containing receptor (NLR) signaling pathways Defensins Fcgamma receptor (FCGR) dependent phagocytosis
	Cytokine signaling in Immune system		Interferon Signaling Signaling by Interleukins Growth Hormone Receptor Signaling Prolactin Receptor Signaling TNFRSF mediated non-canonical NF-kB Pathway ROS, RNS Production in Response to Bacteria
Cell Adhesion	Extracellular matrix organization	Collagen formation	Collagen biosynthesis and modifying enzymes Assembly of collagen fibrils and other multimeric structures Fibronectin matrix formation Elastic fibre formation Laminin interactions Non-integrin membrane-ECM interactions ECM proteoglycans
		Degradation of the extracellular matrix	Activation of Matrix Metalloproteases Collagen Degradation Integrin cell surface interactions
	Cell-Cell communication		Cell junction organization Nephrin interactions
	Vesicle-mediated transport	*Membrane trafficking*	Gap junction trafficking and regulation
Stress	*Cellular responses to stress*		Cellular responses to hypoxia Detoxification of Reactive Oxigen Species

Tab. 4 - List of nodes removed due to excessive interconnection and expected low impact for the scientific question. Nodes were detected in and removed from Reactome's pathways 'RAF/MAP kinase cascade', 'Signaling by the B Cell Receptor (BCR)', 'Immunoregulatory interactions between lymphoid and non-lymphoid cells', 'Complement cascade' and 'Fcgamma receptor (FCGR) dependent phagocytosis'.

Class	Gene symbols
Immunoglobulin heavy chains	IGHD, IGHE, IGHG[1-4], IGHM, IGHV, IGHV3-23, IGHV7-81
Immunoglobulin kappa constant chains	IGKC, IGKV1D-16, IGKV4-1, IGKVA18
Immunoglobulin lambda constant chains	IGLC[1-7]
Immunoglobulin lambda variable chains	IGLV, IGLV1-36, IGLV1-40, IGLV1-44, IGLV10-54, IGLV11-55, IGLV2-11, IGLV2-18, IGLV2-23, IGLV2-33, IGLV3-12, IGLV3-16, IGLV3-22, IGLV3-25, IGLV3-27, IGLV4-3, IGLV4-60, IGLV4-69, IGLV5-37, IGLV5-45, IGLV7-43, IGLV7-46, IGLV8-61
Ig [heavy lambda kappa] chain […] region [XYZ]	multiple (117)
Proteasome complex subunits	PSMA[1-8], PSMB[1-11], PSMC[1-6], PSMD[1-14], PSME[1-4], PSMF1

In order to map the relations between cluster modules and biological pathways, the MSM's node table was exported, reduced to those columns which indicate for name as well as membership to a module and pathways, respectively ('Node_assignment_table_MSM.tsv'). After re-formatting the column names for improved readability, node and edge files were generated by applying two custom scripts (Box 11).

Box 11 - Conversion of MSM node table contents to nodes and edges for a new graph associating modules with pathways.

```
$ awk -v FS=$'\t' \
      -f node_builder.awk \
      Node_assignment_table_MSM_formatted.tsv \
      >> nodes_formatted.tsv
$ awk -v FS=$'\t' \
      -f edge_builder.awk \
      Node_assignment_table_MSM_formatted.tsv \
      >> edges_formatted.tsv
```

In a new Cytoscape session ('modules2pathways.cys'), edges where imported, making up the network. Node attributes were imported afterwards. Colors and sizes of both nodes and edges were set according to the count of underlying MSM nodes and their memberships, respectively. In a session copy ('modules2pathways_max_reduced_allegro.cys'), nodes representing (super-)sets were removed, improving readability. Finally, the Allegro layout "spring electric" was applied with default parameters.

2.7.2 Populating the MSM

As nodes in the MSM represent genes/proteins, all additional information has to be referred to the HGNC approved symbol. As this has been done in the post-processing of the detected imbalanced variants (Box 10), data could be easily appended as table import, matching the 'approved_symbol' column from the data table with the 'shared name' column from Cytoscape's Node Table panel. This step was performed for each of the variant tables resulting from the performed GEMINI queries (Box 6 to Box 9), leading to the populated MSM versions.

Annotations not matching a node in the MSM are not used, in turn nodes with no additional information from the imported data table receive empty fields for the added columns. Consequently, the variant lists are filtered for entries which are not annotated to affect genes of interest (means: the MSM) and visualized in the functional context at the same time.

3. Results

The following sections describe first the MSM specifically generated for the SR usecase, subsequently the sets of variants called from NGS data and their processing, finally the population of the MSM with the filtered candidate variants.

3.1 Description of the Molecular Systems Map

At the end of the process of data from Reactome [Croft *et al.* 2014], the MSM[12] (Fig. 8) included 2,261 nodes (genes/gene products) and 37,769 edges (functional interactions). This refers to 18.57% of 12,177 nodes and 16.47% of 229,300 edges in Reactome overall.

By applying the ReactomeFI clustering (spectral partition based network clustering; [Newman 2006]), 55 modules with between 2 and 420 member genes each were generated. 16 of them showed a size of 24 genes or more ("major modules"). For approving wether those modules are biologically meaningful, a mapping to Reactome's biological pathways was plotted as shown in Fig. 9. Using this mapping, the 16 auto-generated modules with most member nodes could be labeled with biology-oriented descriptions as listed in Tab. 5.

Taken together, genes associated with cell adhesion (collagen, integrin, extracellular matrix; module 1, green) are concentrated in the MSM's lower left region, while WNT-associated cell adhesion mechanisms are separately located in the upper right. A series of dense clusters form a tight half ring around the core region (remaining open at its bottom side).

Most immunology-related mechanisms can be found (e.g. MHC class I (module 2, violet) and class II (module 3, olive)), but also GPCR signaling (module 5, yellow) and nuclear processes (module 6, pink; e.g. pore complex). In the periphery regions, several receptor-related upstream mechanisms are located, like TGFβ signaling (module 10, maroon) and WNT (modules 8, 11, 13 in dark green, green blue and light pink, respectively), but also the complement cascade (module 15).

Interestingly, EGFR as a (semantically) central molecule in this use case, was located right in the center of the MSM's layout. Here, in the core region, most genes (including EGFR) are accounted for the biggest module 0 (blue), making up the signaling core of most pathways. Signaling includes shared components like PIP3, AKT1 and MAP kinases, but also major parts of e.g. the FGFR, VEGF and the other ERBB pathways.

[12] For a high-resolution print, please refer to the foldable inlay map at the end of this thesis. For a zoomable vector graphic, please refer to the digital contents.

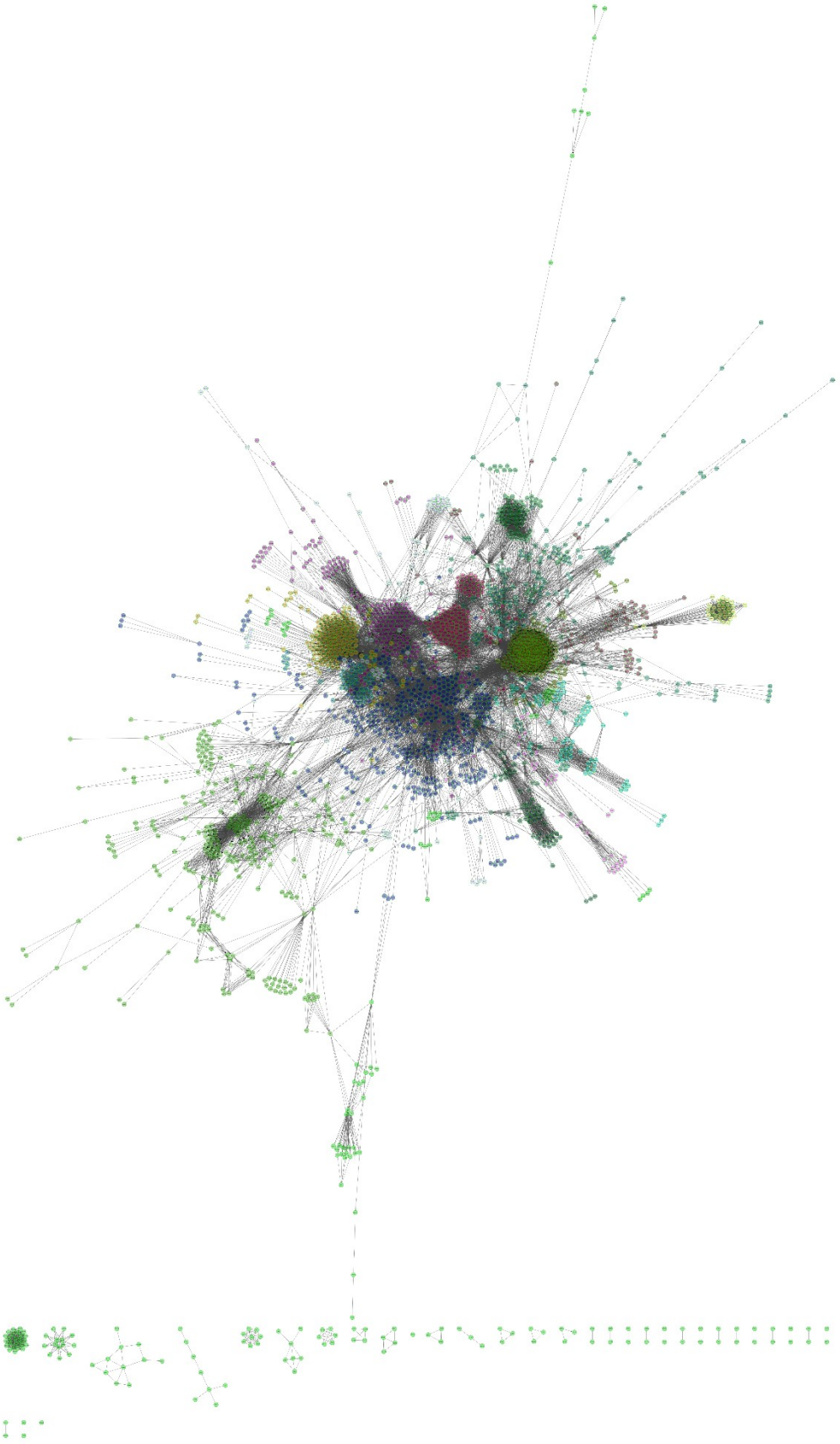
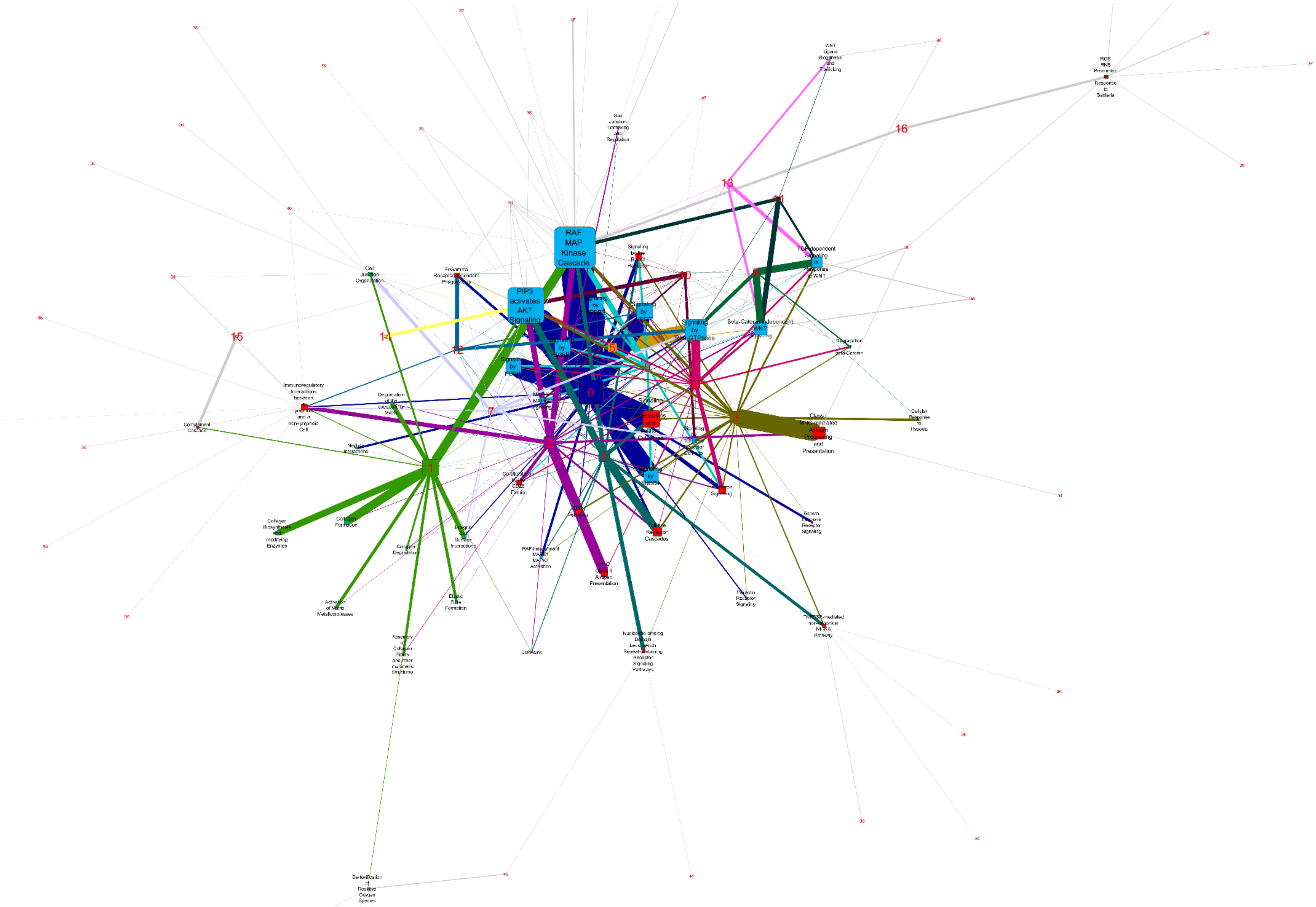


Fig. 8 - Blank Molecular Systems Map. ReactomeFI's clustering algorithm calculated functional modules (indicated by colors for the 14 modules with most members), while layout was performed by Allegro. Small clusters and single nodes unconnected to the majority of nodes are sorted to the bottom of the map.

Tab. 5 - Description of clustered modules according to pathway memberships of included nodes. Restriction to the 16 modules with most member nodes ("major modules"), colors following Fig. 8 and Fig. 9 (auto-generated by Cytoscape).

Module ID	# Nodes	Color	Description
0	420		Growth factor signaling, cytokines
1	227		Extracellular matrix mechanisms
2	208		MHC class II signaling (adaptive immunity, extracellular)
3	203		MHC class I signaling (adaptive immunity, intracellular)
4	189		TLR/Interleukin/TNF/LRR signaling (innate immunity, extracellular)
5	149		Rho-GTPases/GPCR signaling
6	127		Nucleus mechanics (nuclear pore complex, chromatin)
7	89		Cell interaction/junction upon Rho/TGF beta
8	75		WNT signaling
9	73		EGFR/ERBB/FGFR downstream signaling (e.g. mTOR)
10	69		TGF beta signaling
11	63		WNT signaling (downstream)
12	46		Immunoregulatory phagocytosis (Fc Gamma)
13	35		WNT signaling (biogenesis/trafficking)
14	32		DNA replication, cell cycle
15	30		Complement Cascade
16	24		ROS/RNS and signaling

Fig. 9 - Mapping of clustering modules to biological pathways (next page). Numbered boxes represent MSM modules clustered automatically by ReactomeFI plugin, while all other boxes represent biological pathways. Genes of the MSM can be member of multiple pathways, but only one module. Layout has been performed by applying Allegro with edge weighting. The box size linearly depends on member count for modules or pathways. Width of connecting edges indicates for the number of genes shared by adjacent boxes. Color codes for modules 1-14 are used as in Fig. 8, other modules are kept in grey. Colors for pathway boxes are chosen by membership to a superset (blue = signal transduction, red = immune system, green = cell adhesion, grey = stress). Edge coloring follows the adjacent module node. Full graphical details available in the original image file (supplementary materials).



RAF MAP Kinase Cascade
PIP3 activates AKT Signaling
Signaling by Rho GTPases
WNT Ligand Biogenesis and Trafficking
ROS RNS Production Response to Bacteria
TCF dependent Signaling in Response to WNT
Beta-Catenin independent WNT Signaling
Degradation of Beta-Catenin
Cellular Response to Hypoxia
Class I MHC-mediated Antigen Processing and Presentation
Immunoregulatory Interactions between a Lymphoid and a non-Lymphoid Cell
Complement Cascade
Cell Junction Organization
FcGamma Receptor dependent Phagocytosis
Degradation of the extracellular Matrix
Collagen Biosynthesis and modifying Enzymes
Collagen Formation
Collagen Degradation
Activation of Matrix Metalloproteinases
Elastic Fibre Formation
Assembly of Collagen Fibrils and other multimeric Structures
Detoxification of Reactive Oxygen Species
Defensins
RAF-independent MAPK1 MAPK3 Activation
MHC Class II Antigen Presentation
Nucleotide-binding Domain Leucine-rich Repeat-containing Receptor Signaling Pathways
Interleukin Receptor Signaling
TNFR2 non-canonical NF-kB Pathway
Growth Hormone Receptor Signaling
Gap Junction Trafficking and Regulation
0
1
2
3
4
5
6
7
10
11
12
13
14
15
16

3.2 NGS quality control: raw reads, mapping and variant calling

Starting with the given input data, quality was checked in major steps of the processing pipeline. Raw reads Kmer contents were indicated by FastQC to be suboptimal due to homopolymeric sequences of A and T and showed a slight, expected drop of quality from 5’ to 3’ end. After quality-based trimming on both ends, empty reads remained, triggering QC errors for over-represented sequences (empty string) and a suspicious sequence length distribution.

After mapping the reads to the reference genome, additional warnings for GC content were reported. Integrating read re-ordering, duplicate marking and re-alignment fixed all issues, with a remaining warning of sequence length distribution referring to kept empty reads. Keeping empty reads appeared to be necessary for compatibility reasons in later stages of the pipeline, as some tools cannot cope with missing pair members in paired-end datasets.

Regarding the mapping itself, FastQC reported no critical issues, although some parameters on qualities (base and sequence level) were marked as being not optimal. An exception is the dataset of patient 072, where quality values were noted as being insufficient, mainly due to average PHRED scores around 18 for both per-base and per-sequence statistics. The other datasets show values around 25, with 20 being a common lower limit. However, as this particular patient is of the group of interest (SR-negative), which is anyhow smaller than the SR-positive one, and coverage was highest among the whole cohort, the patient’s dataset was kept for subsequent analyses.

For gaining insights in comparability of patient’s samples and the results of the three applied variant callers, a visual inspection was introduced by systematically querying the loaded GEMINI database and plotting the results (comp. chapter 2.5). Except for dataset 072, the X-Y scatterplots appeared uniformly for any chromosome when comparing various attribute combinations (for examples refer to Figs. Fig. 10 to Fig. 12, for the overview to the digital supplements). Regarding dataset 072, two major observations were made: called variants are much more dedicated to exonic positions, and the quality value distributions do not show a difference to the other datasets (Fig. 13), supporting the decision to work further on with these data. Taken together, the raw data and their processings were considered to be valid.

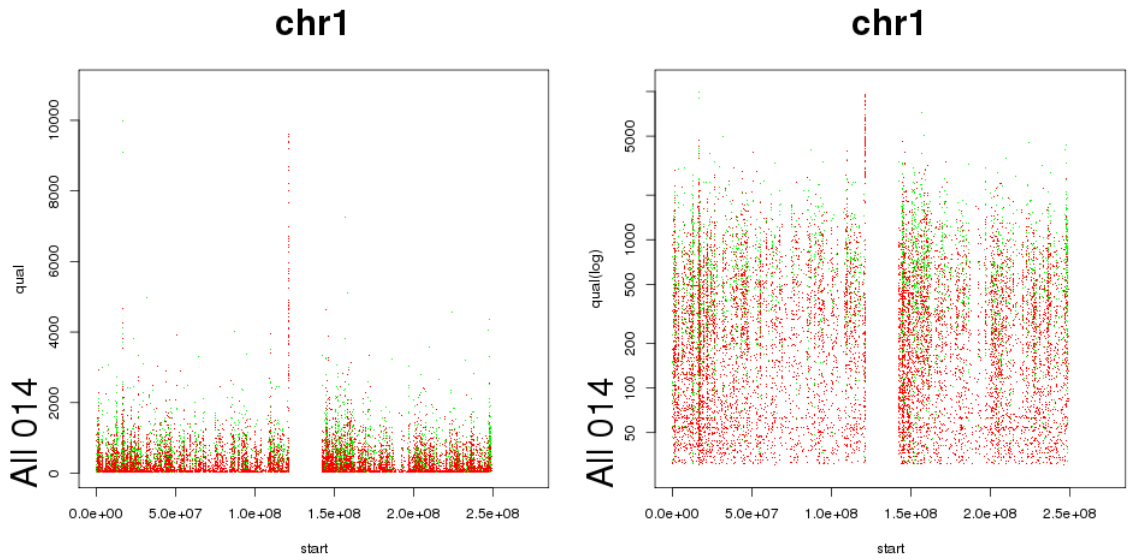


Fig. 10 - Genomic position vs. VCF quality ('QUAL') example plot. Qualities of all called variants, independent of any filtering, were plotted according to their start position (VCF 'POS') on a particular chromosome for one patient each (here: chr1 for individual 014) for visual inspection. The VCF 'QUAL' column averages all sample measurements here: (sample = variant caller). Exonic calls are obviously of higher quality in average. Red = intronic/intragenic call, green = exonic call. The gap in the middle marks the centromer. Left: linear scales; right: log scale for quality.

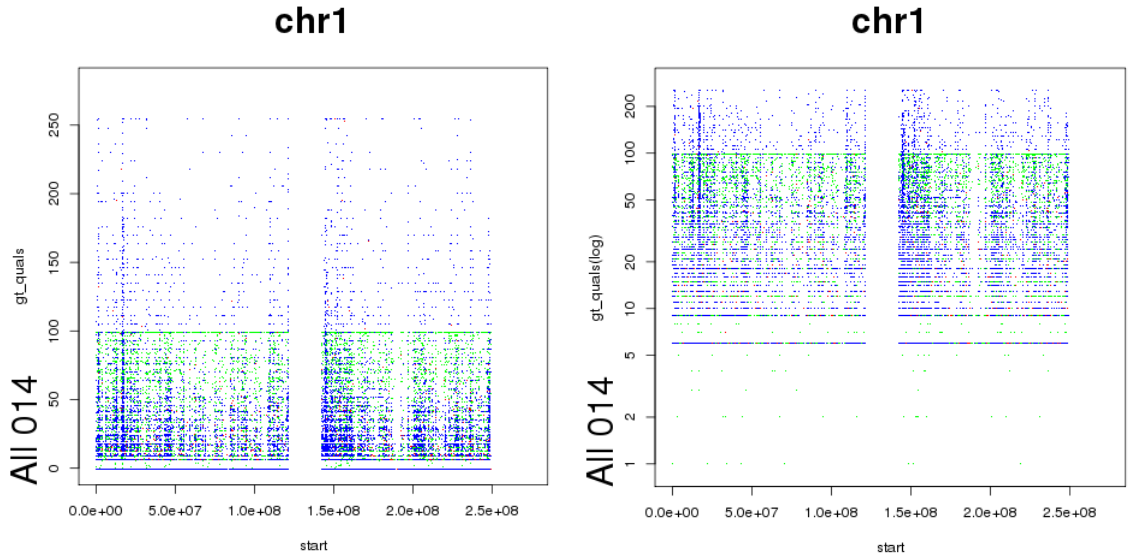


Fig. 11 - Genomic position vs. variant caller quality ('GQ') example plot. Qualities of all called variants, independent of any filtering, were plotted according to their start position (VCF 'POS') on a particular chromosome for one patient each (here: chr1 for individual 014) for visual inspection. The VCF sample columns carry a field 'GQ' ('gt_quals' in GEMINI) indicating for a non-averaged quality value dedicated to a particular (here: sample = variant caller). Blue = GATK Haplotype Caller, green = VarScan SNP, red = VarScan InDel. The gap in the middle marks the centromer. Left: linear scales; right: log scale for quality.

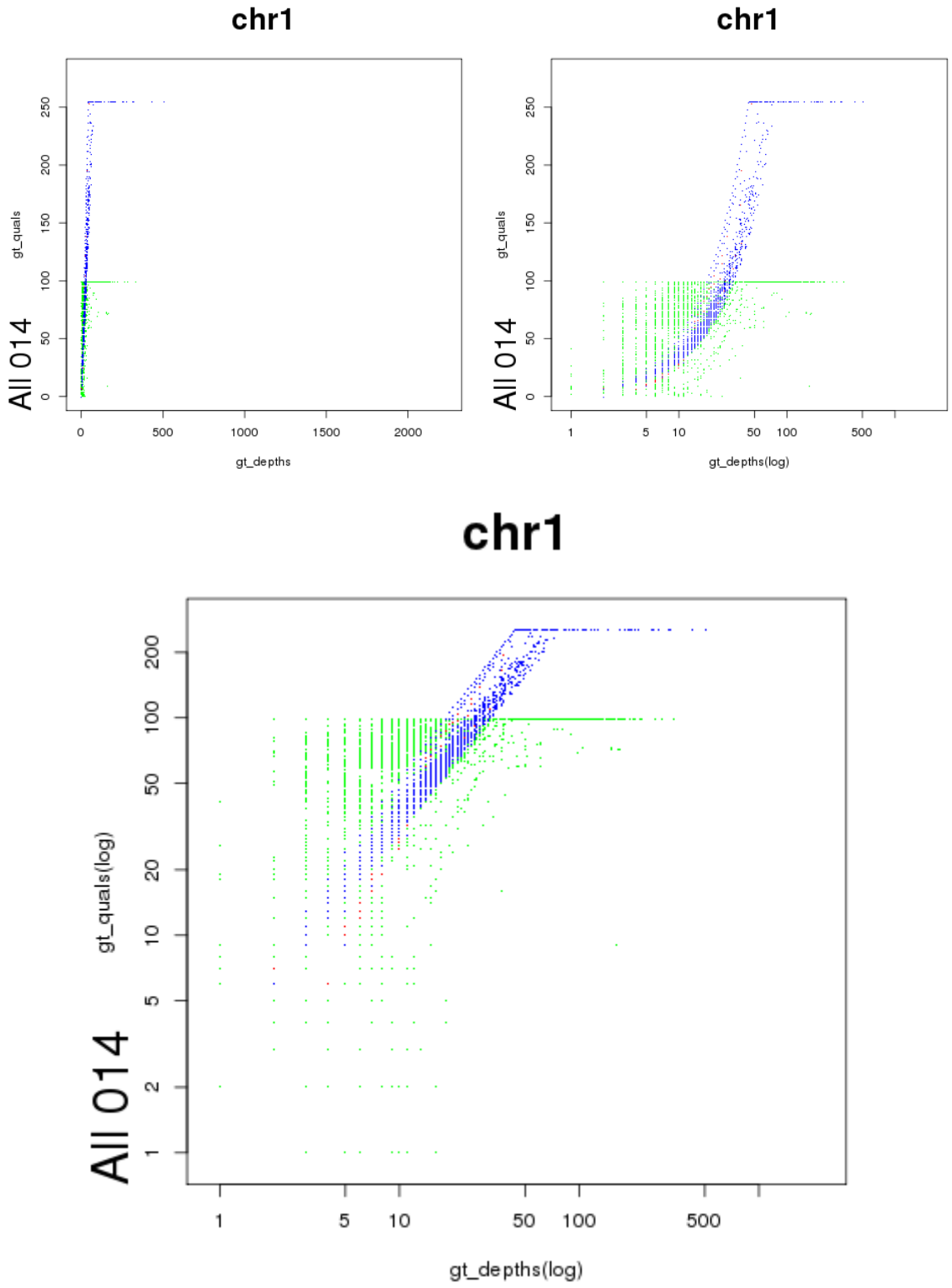


Fig. 12 - Variant coverage vs. variant caller quality ('GQ') example plot. Qualities of all called variants, independent of any filtering, were plotted against according to their start position (VCF 'POS') on a particular chromosome for one patient each (here: chr1 for individual 014) for visual inspection. The VCF sample columns carry a field 'GQ' ('gt_quals' in GEMINI) indicating for a non-averaged quality value dedicated to a particular (here: sample = variant caller). Both tools hit their maxima (GATK: 255; VarScan: 100) with several variants. Blue = GATK Haplotype Caller, green = VarScan SNP, red = VarScan InDel. Top left: linear scales; top right: log scale for quality; bottom: log scale for both.

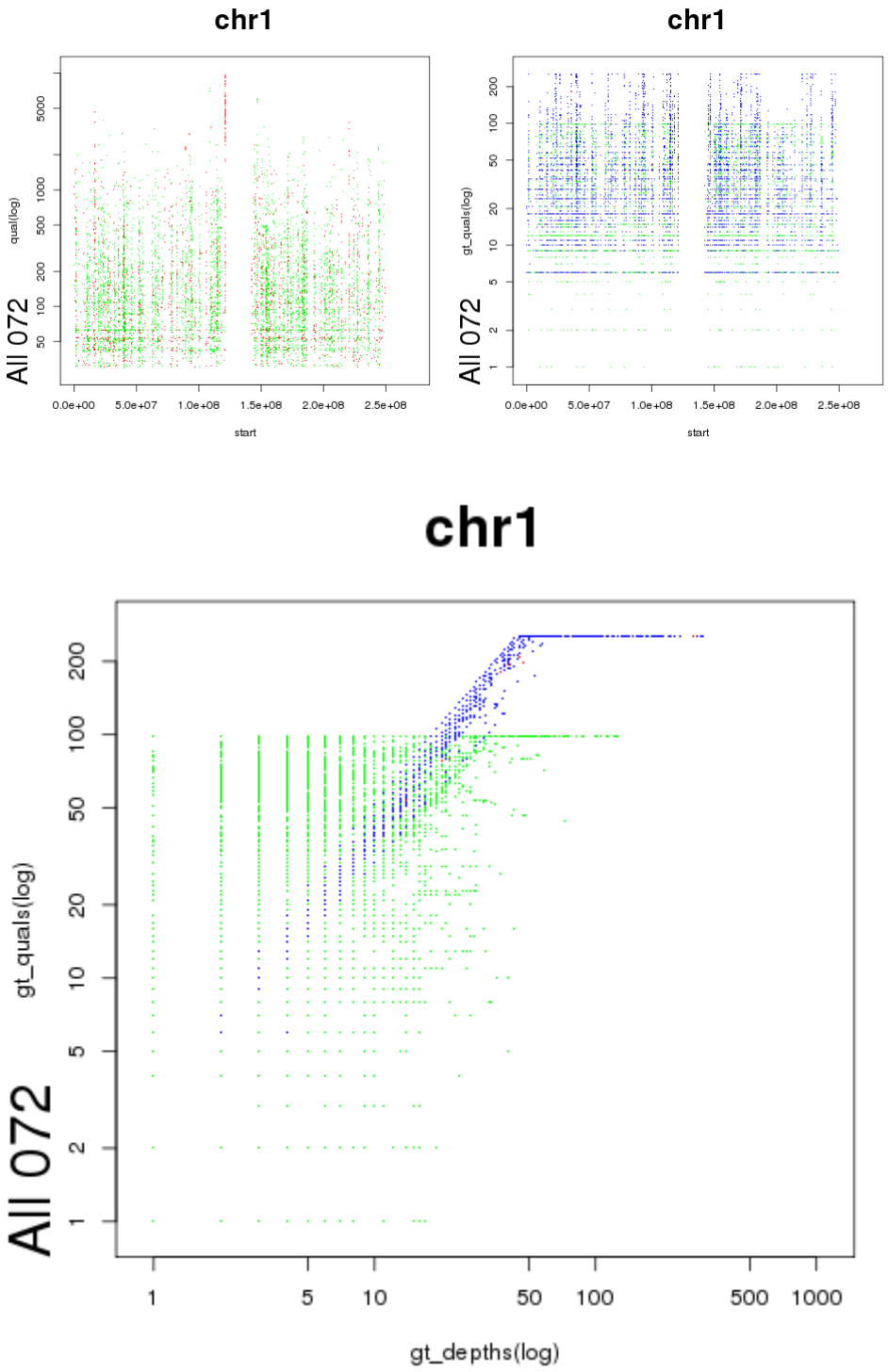


Fig. 13 - Plots of outlying dataset 072. Ranges and distributions appear in all three perspectives like they were found in other datasets (top left: comp. Fig. 10, right panel; top right: comp. Fig. 11, right panel; bottom: comp. Fig. 12, bottom panel). Exception is the low number of intronic variants and a consequently low density overall.

3.3 Filter assessment and imbalance detection

Principally, both applied filtering mechanisms (SNPeff impact class and zygosity of a variant) should be independent. In order to estimate both their effects, the sets defined by the 2x2 cross-table presented in Tab. 6 were analyzed for their overall filtering on those variants passing both the coverage (>= 5x) and quality/confidence (Phred >= 13; ≙ 95%) filter in at least one sample, without any restriction on imbalance criteria ('baseline').

Tab. 6 - Filtering effects on baseline variants according to impact class and zygosity (cross-table). While ignoring heterozygously called variants cuts down the list by ~60%, filtering for an impact class of 'MED' or 'HIGH' leaves just 10% of the initially called variants. These numbers hold relatively independently of whether the other filter is set or not. The combined application (Set 4) yields in just 4.2% of the initial Set 1.

Zygosity (genotype) → cutting ~ 60%	Impact strength (SNPeff) → cutting ~90%	
	any	no LOW
any	Set 1: **514,099** variants	Set 3: **57,869** variants
homozygous	Set 2: **219,182** variants	Set 4: **21,540** variants

In the actual GEMINI calls (filtering by imbalance >= 7) for sure the number of remaining variants dropped dramatically, but in parallel for all three sets on a logarithmic scale (comp. Fig. 14).

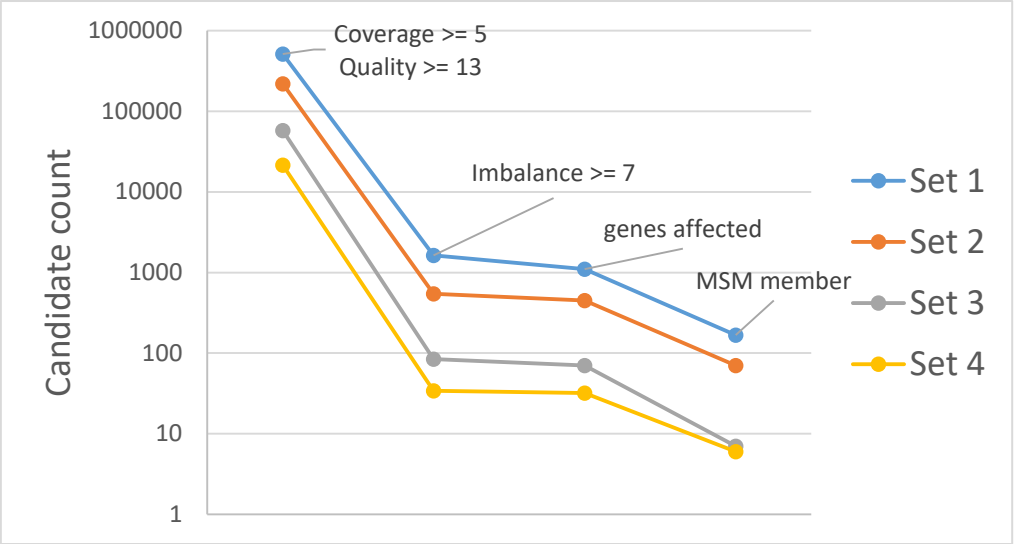


Fig. 14 - Candidate count according to filtering steps applied in the order of application to all four sets. Sets according to Tab. 6. X axis reads as progress in data processing. Labels in columns define filters applied, resulting in the depicted number of remaining candidates. Columns one and two indicate for variants, three and four to affected genes. Filtering effects obviously apply homogeneously to all sets.

Subsequently, using the approved HGNC gene names, the four separate variant sets were each converted to gene-centric sets (comp. Tab. 6). Those could then be assigned to a copy of the MSM each (comp. Section 3.4).

While querying the database for significantly imbalanced variants resulted in 1,629 variants overall, 545 of those were reported to be homozygous (Tab. 7). The exclusion of variants classified as being low impact by SNPeff reduced the number of candidate variants even more (84 with 34 being homozygous). Low impact filtering excludes all intergenic and intronic variants. Applying impact and/or zygosity filters did not change the numerical relation to affected transcripts or genes, as indicated by a high R^2 for both genetic elements (Fig. 15).

Tab. 7 - Sets of detected imbalanced variants across all patients. Relation of variants to affected functional genetic elements (transcripts and genes) according to SNPeff annotation. Low impact variants are mainly intronic. HOM = homozygous, HET= heterozygous.

Set	genotype	LOW impact	initial variant count	imbalanced variants	affected transcripts	affected genes
1	All (HOM + HET)	included	514,099	1,629	10,578	1,104
2	HOM only	included	219,182	545	4,350	448
3	All (HOM + HET)	excluded	57,869	84	571	70
4	HOM only	excluded	21,540	34	272	32

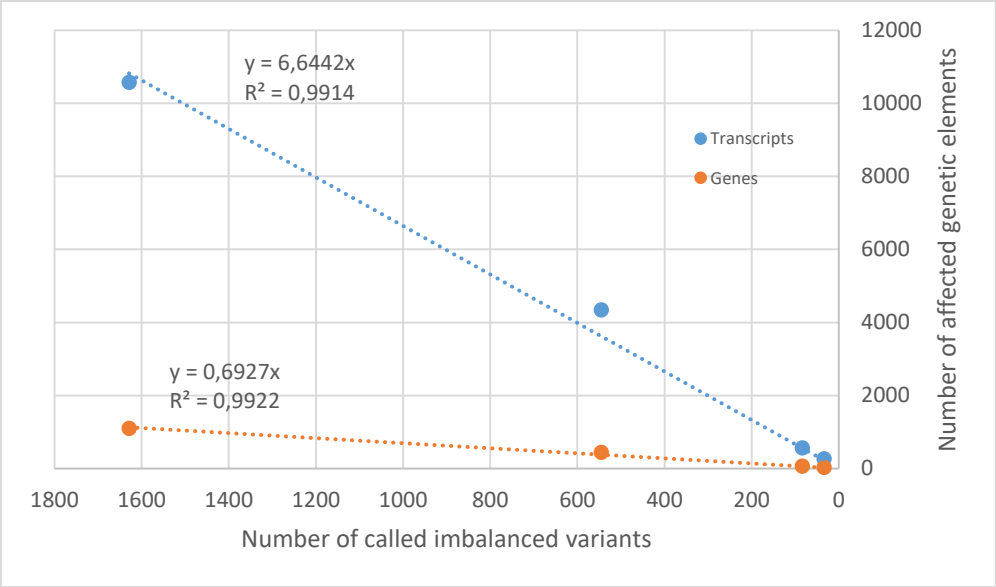


Fig. 15 - Imbalanced variants vs. affected genetic elements. Count relations and linear fit indicate no principal change by applying low impact or zygosity filtering. Values from Tab. 5. Data points correspond to „ All (HOM + HET)“, „ HOM only“, „All (HOM + HET), LOW impact excluded“ and „HOM only, LOW impact excluded”, respectively, read from left to right.

In terms of patient assignments, the imbalanced variants appeared more often in the SR-negative individuals (Tab. 8). This refers to the higher number of imbalanced variants assigned to the SR-positive group, a consequence of the unequal group sizes (11x negative, 12x positive).

Tab. 8 - Imbalanced variants accounted for individual patients. Again, 072 appears as an outlier.

SR-negative patients	imbalanced variants	SR-positive patients	imbalanced variants
036	614	**014**	663
072	12	**111**	1317
125	307	**137**	1245
020	226	**155**	1327
281	363	**090**	1342
375	204	**213**	1302
406	190	**344**	1298
428	327	**566**	1341
586	631	**598**	1363
750	602	**624**	1171
796	502	**638**	1360
		708	1323

Runtimes were unexpectedly long; GEMINI is a database system, although implemented in SQLite, so SELECT statements of the given complexity should not take hours. Even when not formulating the explicit JOIN on `variants` and `gene_detailed` table, a substantial speed-up could not be observed, so GEMINI remained a bottleneck. Finally, the generation of Sets 1 and 3 took 8h 47min and 8h 45min, respectively. For the more restricted Sets 2 and 4, in which low impact variants are already excluded with the database call, query time was at 1h 38min and 1h 45min, respectively.

When projecting the called variants onto the statistical model, distributions in Fig. 16 (all variants) and Fig. 17 (homozygeous variants only) show the relation between minor allele frequency (MAF according to 1000 Genomes Project [Auton *et al.* 2015]) and the probability to occur in the measured imbalanced manner in the given patient cohort, following the model. In both plots, those variants being specifically detected for the patient group of interest (SR-negative) where highlighted in order to compare visually the distribution of this sub-population.

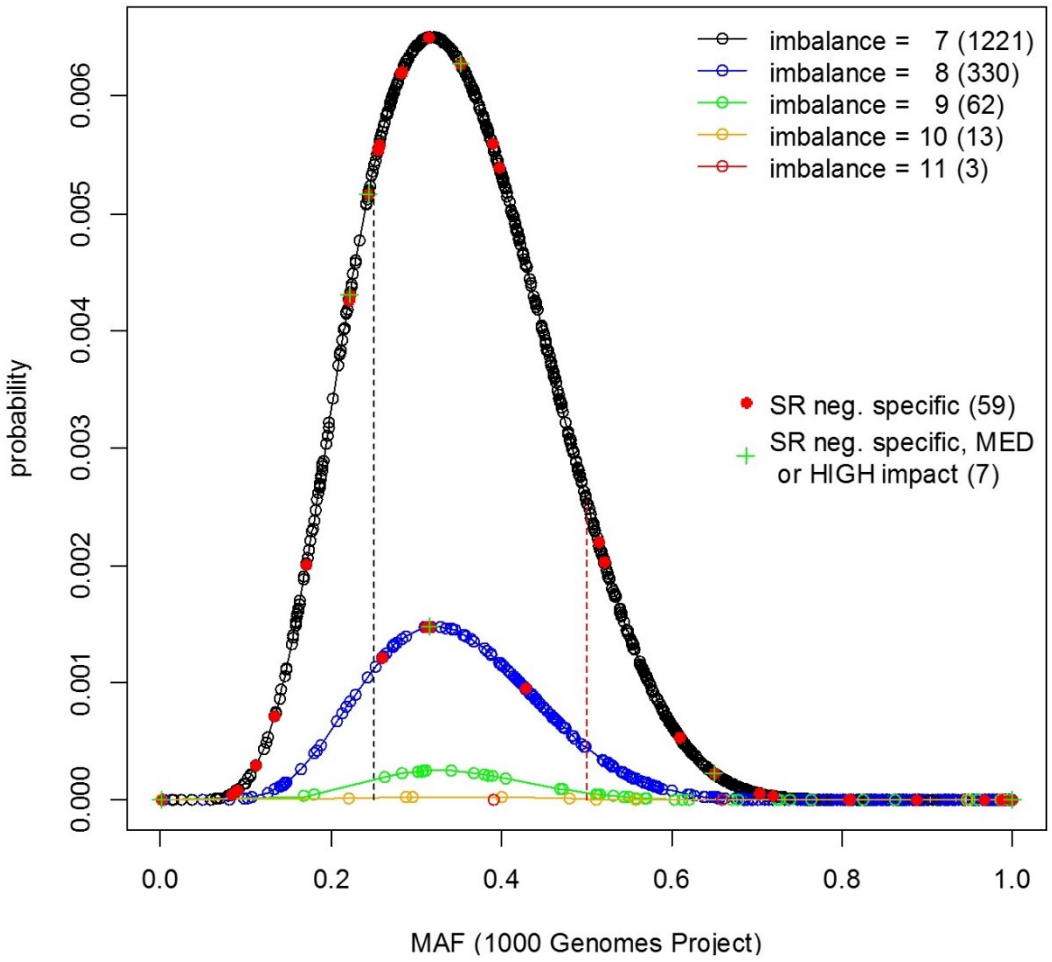


Fig. 16 - Distribution of called variants according to the statistical model of imbalance. 48 out of 59 variants being specifically overrepresented in the SR-positive group show a minor allele frequency (MAF) of 0.5 or lower (red dashed vertical line), 36 out of 59 variants are rated with a MAF of 0.25 or lower (black vertical dashed line). Focussing to the 7 of those SR-positive specific variants, which are annotated to be medium (MED) or high (HIGH) impact by SNPeff, 5 or 2 of them are listed with a MAF of 0.5 or lower and 0.25 or lower, respectively. Lines underlying the points of the four imbalance classes with >= 4 detections are splines smoothing the distribution.

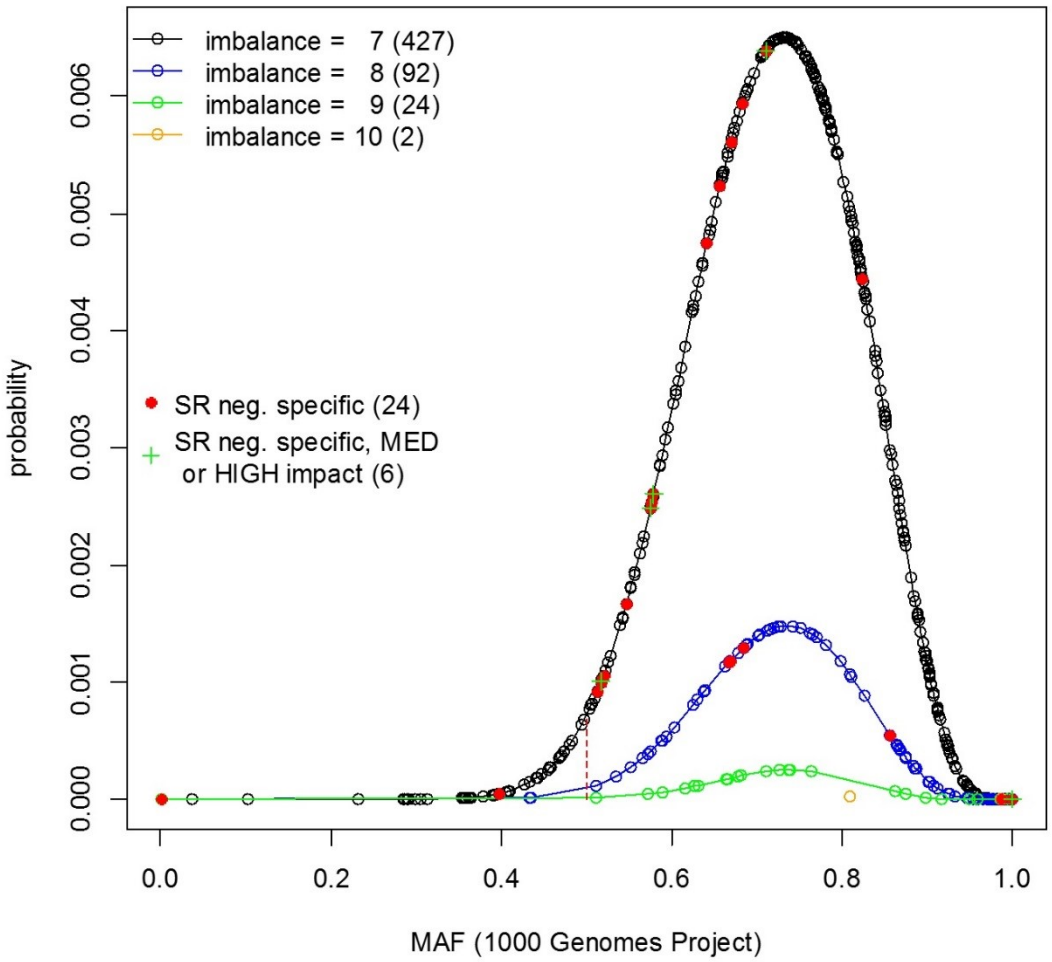


Fig. 17 - Distribution of called homozygous variants only, according to the statistical model of imbalance. 3 out of 24 variants being specifically overrepresented in the SR-positive group show a minor allele frequency (MAF) of 0.5 or lower (red dashed vertical line), 1 out of 24 variants is rated with a MAF of lower than 0.25. Focussing to the 6 of those SR-positive specific variants, which are annotated to be medium (MED) or high (HIGH) impact by SNPeff, none of them is listed with a MAF below 0.5. Lines underlying the points of the three imbalance classes with >= 4 detections are splines smoothing the distribution.

3.4 Joining imbalanced variants with the MSM: populating the map

After having mapped the called and filtered variants to the genes they are predicted to affect, those genes could be located within the MSM if anyhow present. Consequently, the MSM's filtering function was used in order to finally reduce the remaining candidates to those being of potential mechanistic interest.

In average, 15% of the affected genes were assigned to be part of the MSM, while around 18% of all known genes are represented in Reactome's database. Simultaneously, the fraction of MSM-located genes specificially hit in the SR-negative group increases with the degree of filtering from less than 2% up to 50% (comp. Tab. 11).

Tab. 9 - Variants detected to occur imbalanced specifically in the SR-negative group. Sorting by genomic position. All impacts on multiple transcripts were predicted to be equal in terms of type and severity. Entries printed in bold are part of the MSM (see Section 2.7).

chrom	Start [bp]	Ref. allele	Alt. allele	rs IDs [dbSNP]	MAF [1kg project]	Impact type [Sequence Ontology (SO)]	Impact severity [SO]	SR neg. count	SR pos. count	diff	Exclusive for SR neg. group?	Affected gene	Number of affected transcripts
chr1	43305752	G	A	rs12034000	0.260	intron_variant	LOW	8	0	8	YES	ERMAP	5
chr1	204159607	GC	G	-	-	frameshift_variant	HIGH	8	1	7		KISS1	1
chr1	204159610	CT	C	rs71745629	0.222	frameshift_variant	HIGH	8	1	7		KISS1	1
chr1	216824504	C	G	rs10863261	0.134	intron_variant	LOW	8	1	7		ESRRG	23
chr2	73339707	G	A	rs1864488	0.521	synonymous_variant	LOW	9	2	7		RAB11FIP5	6
chr2	107460473	T	G	rs12473361	0.256	5_prime_UTR_variant	LOW	10	3	7		ST6GAL2	5
chr2	239010718	T	C	rs36066915	0.609	synonymous_variant	LOW	7	0	7	YES	ESPNL	5
chr4	983808	C	T	rs4690221	0.389	synonymous_variant	LOW	7	0	7	YES	SLC26A1	4
chr6	13010090	TG	T	-	-	intron_variant	LOW	7	0	7	YES	PHACTR1	13
chr6	13010091	GCCA	G	rs142689950	0.967	intron_variant	LOW	7	0	7	YES	PHACTR1	13
chr6	13010091	GCCACCA	G	-	-	intron_variant	LOW	7	0	7	YES	PHACTR1	13
chr6	44151764	G	A	rs13209117,rs386432577	0.311	3_prime_UTR_variant	LOW	8	0	8	YES	CAPN11	6
chr7	26678855	A	G	rs57000259	0.397	synonymous_variant	LOW	8	1	7		C7orf71	1
chr7	30634660	C	G	rs1049402	0.649	missense_variant	MED	7	0	7	YES	GARS	9
chr7	122769615	T	C	rs13243580	0.221	intron_variant	LOW	8	1	7		SLC13A1	4
chr7	132248810	C	CTGT	rs147358337	-	intron_variant	LOW	7	0	7	YES	PLXNA4	5
chr7	152544636	T	C	rs386622381,rs940262	0.808	intron_variant	LOW	8	1	7		ACTR3B	6
chr8	6500543	C	T	rs1057091	0.244	missense_variant	MED	8	1	7		MCPH1	7
chr9	86585678	A	G	-	-	synonymous_variant	LOW	7	0	7	YES	HNRNPK	12
chr9	86585690	A	G	-	-	synonymous_variant	LOW	7	0	7	YES	HNRNPK	12
chr9	86585714	A	G	-	-	synonymous_variant	LOW	7	0	7	YES	HNRNPK	12

chr9	86586983	G	T	-	-	synonymous_variant	LOW	7	0	7	YES	HNRNPK	12
chr9	86586990	A	G	-	-	synonymous_variant	LOW	7	0	7	YES	HNRNPK	12
chr9	86586999	T	A	-	-	synonymous_variant	LOW	7	0	7	YES	HNRNPK	12
chr9	86587002	G	T	-	-	synonymous_variant	LOW	7	0	7	YES	HNRNPK	12
chr9	86587008	G	T	-	-	synonymous_variant	LOW	7	0	7	YES	HNRNPK	12
chr9	86587059	G	A	rs200783060	-	synonymous_variant	LOW	7	0	7	YES	HNRNPK	12
chr9	86587065	A	G	-	-	synonymous_variant	LOW	7	0	7	YES	HNRNPK	12
chr9	86587080	C	T	-	-	synonymous_variant	LOW	7	0	7	YES	HNRNPK	12
chr9	126128252	G	C	rs1105222	0.316	missense_variant	MED	8	0	8	YES	CRB2	4
chr11	1093451	G	A	rs34136803	-	synonymous_variant	LOW	9	2	7		MUC2	4
chr11	2869187	C	T	rs11601907	0.083	synonymous_variant	LOW	8	1	7		KCNQ1	6
chr11	45230861	C	A	rs3781705	0.255	intron_variant	LOW	8	1	7		PRDM11	6
chr11	63767185	A	G	rs709594	0.428	synonymous_variant	LOW	10	2	8		MACROD1	8
chr11	63767248	C	T	rs11600062	0.171	intron_variant	LOW	8	1	7		MACROD1	8
chr12	1019943	G	A	rs2023944	0.987	3_prime_UTR_variant	LOW	8	1	7		WNK1	21
chr12	**6938022**	**C**	**CG**	**-**	**1.000**	**frameshift_variant**	**HIGH**	**7**	**0**	**7**	**YES**	**P3H3**	**13**
chr13	113158859	AC	A	-	-	intron_variant	LOW	8	1	7		TUBGCP3	6
chr13	113158860	CATAT	C	rs3832905,rs397840783	0.282	intron_variant	LOW	8	1	7		TUBGCP3	6
chr13	113158860	CAT	C	-	-	intron_variant	LOW	8	1	7		TUBGCP3	6
chr13	113158860	C	CAT	-	-	intron_variant	LOW	8	1	7		TUBGCP3	6
chr14	**103879023**	**G**	**A**	**rs4900574**	**0.703**	**intron_variant**	**LOW**	**9**	**2**	**7**		**MARK3**	**26**
chr15	28090172	C	T	rs12592307	0.283	synonymous_variant	LOW	7	0	7	YES	OCA2	5
chr16	1393019	G	A	rs2235632	0.352	splice_region_variant	MED	10	3	7		BAIAP3	16
chr17	41465914	GTC	G	-	-	exon_variant	LOW	7	0	7	YES	LINC00910	7
chr17	41465918	CAAGATAT	C	-	-	exon_variant	LOW	7	0	7	YES	LINC00910	7
chr17	41465926	CATCA	C	-	-	exon_variant	LOW	7	0	7	YES	LINC00910	7

chr17	41465936	G	GCA	-	-	exon_variant	LOW	7	0	7	YES	LINC00910	7
chr17	41465938	A	G	-	-	exon_variant	LOW	7	0	7	YES	LINC00910	7
chr17	41465940	A	G	-	-	exon_variant	LOW	7	0	7	YES	LINC00910	7
chr17	41465941	C	CG	-	-	exon_variant	LOW	7	0	7	YES	LINC00910	7
chr17	41465943	A	G	-	-	exon_variant	LOW	7	0	7	YES	LINC00910	7
chr17	41465949	A	C	rs140897052	-	exon_variant	LOW	7	0	7	YES	LINC00910	7
chr17	80275252	T	C	rs3176829	0.887	intron_variant	LOW	7	0	7	YES	CD7	7
chr19	58898192	C	T	rs975947	0.314	exon_variant	LOW	7	0	7	YES	MIR4754	1
chr20	2594189	C	T	rs10485601	0.090	intron_variant	LOW	7	0	7	YES	TMC2	2
chr20	62318855	G	C	rs6062490	0.718	intron_variant	LOW	7	0	7	YES	RTEL1	11
chr22	36623147	T	C	rs132759	0.514	3_prime_UTR_variant	LOW	8	1	7		APOL2	9
chrX	**41530654**	**A**	**AT**	**rs199561120**	**0.112**	**intron_variant**	**LOW**	**7**	**0**	**7**	**YES**	**CASK**	**15**

Tab. 10 - Homozygous variants detected to occur imbalanced specifically in the SR-negative group. Sorting by genomic position. All impacts on multiple transcripts were predicted to be equal in terms of type and severity. Entries printed in bold are part of the MSM (see Section 2.7).

chrom	Start [bp]	Ref. allele	Alt. allele	rs IDs [dbSNP]	MAF [1kg project]	Impact type [Sequence Ontology (SO)]	Impact severity [SO]	SR neg. count	SR pos. count	diff	Exclusive for SR neg. group?	Affected gene	Number of affected transcripts
chr1	26310393	A	G	rs2275102,rs386486547	0.547	intron_variant	LOW	9	2	7		PAFAH2	7
chr1	72740633	G	A	rs7549792	0.397	intron_variant	LOW	7	0	7	YES	NEGR1	6
chr2	54122955	A	G	rs698856	0.823	intron_variant	LOW	7	0	7	YES	PSME4	10
chr3	**13368891**	**G**	**A**	**rs2271509**	**0.396**	**synonymous_variant**	**LOW**	**7**	**0**	**7**	**YES**	**NUP210**	**4**
chr3	**13383539**	**A**	**G**	**rs2271504**	**0.577**	**synonymous_variant**	**LOW**	**7**	**0**	**7**	**YES**	**NUP210**	**4**
chr3	**13395578**	**C**	**A**	**rs2280084**	**0.575**	**missense_variant**	**MED**	**7**	**0**	**7**	**YES**	**NUP210**	**4**
chr3	**13407555**	**T**	**C**	**rs3732671**	**0.579**	**missense_variant**	**MED**	**7**	**0**	**7**	**YES**	**NUP210**	**4**
chr3	**13415145**	**G**	**T**	**rs6810356**	**0.517**	**intron_variant**	**LOW**	**7**	**0**	**7**	**YES**	**NUP210**	**4**

chr3	**13418085**	**C**	**A**	**rs6442377**	**0.512**	**intron_variant**	**LOW**	**7**	**0**	**7**	**YES**	**NUP210**	**4**
chr3	**13421149**	**C**	**T**	**rs7628051**	**0.517**	**missense_variant**	**MED**	**7**	**0**	**7**	**YES**	**NUP210**	**4**
chr3	**13421319**	**T**	**C**	**rs7625586**	**0.520**	**intron_variant**	**LOW**	**7**	**0**	**7**	**YES**	**NUP210**	**4**
chr3	**14551443**	**A**	**G**	**rs9845816**	**0.710**	**missense_variant**	**MED**	**7**	**0**	**7**	**YES**	**GRIP2**	**7**
chr3	134322741	A	G	rs2293293	0.684	synonymous_variant	LOW	10	2	8		KY	5
chr3	134331979	C	T	rs9838119	0.683	intron_variant	LOW	8	1	7		KY	5
chr5	54830379	A	G	rs11745331	0.856	intron_variant	LOW	8	0	8	YES	PLPP1	6
chr5	**140955776**	**C**	**T**	**rs2302103**	**0.641**	**intron_variant**	**LOW**	**8**	**1**	**7**		**DIAPH1**	**18**
chr5	**141016287**	**T**	**G**	**rs1421896**	**0.656**	**intron_variant**	**LOW**	**8**	**1**	**7**		**HDAC3**	**16**
chr7	5427719	A	G	rs4724663	0.999	missense_variant	MED	9	2	7		TNRC18	10
chr7	16504177	A	T	rs4506094	0.669	intron_variant	LOW	8	1	7		SOSTDC1	2
chr7	132248810	C	CTGT	rs147358337	-	intron_variant	LOW	7	0	7	YES	PLXNA4	5
chr10	134981689	T	C	rs2998152	0.669	intron_variant	LOW	8	0	8	YES	KNDC1	6
chr10	134981836	T	C	rs2998151	0.667	intron_variant	LOW	8	0	8	YES	KNDC1	6
chr12	1019943	G	A	rs2023944	0.987	3_prime_UTR_variant	LOW	8	1	7		WNK1	21
chr12	**6938022**	**C**	**CG**	**None**	**1.000**	**frameshift_variant**	**HIGH**	**7**	**0**	**7**	**YES**	**P3H3**	**13**

Tab. 11 - Filtering effects on gene-centered variant sets by applying the MSM. Percentages for MSM-located genes express the relation to affected genes overall. Percentages of SR-negative genes in turn refer to fractions in MSM-located genes.

Set	genotype	LOW impact	affected genes overall	genes located within MSM	genes assigned to SR-negative group
1	All (HOM + HET)	included	1,104	167 (15.12%)	3 (1.8%)
2	HOM only	included	448	70 (15.63%)	5 (7.1%)
3	All (HOM + HET)	excluded	70	7 (10,00%)	1 (14.3%)
4	HOM only	excluded	32	6 (18.75%)	3 (50.0%)

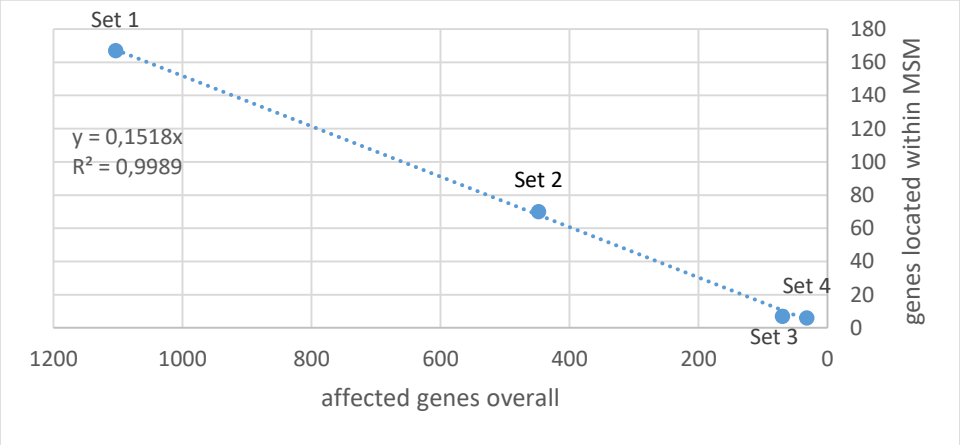


Fig. 18 - Relation of MSM filtering effect across sets. A strong trend shows 15% of affected genes remaining when applying the MSM as a pathway-driven filter on the different sets.

For the maps displaying also LOW impact data (Sets 1 and 2; Fig. 19 and Fig. 20, respectively), overviews were generated in order to get an impression of the topological distribution of affected genes. For both sets, an accumulation of events specific for SR-positive cases could not be observed in any area of the map. Despite the quite low number, for Set 1 events specific for SR-negative cases could only be located in and between core and lower left area.

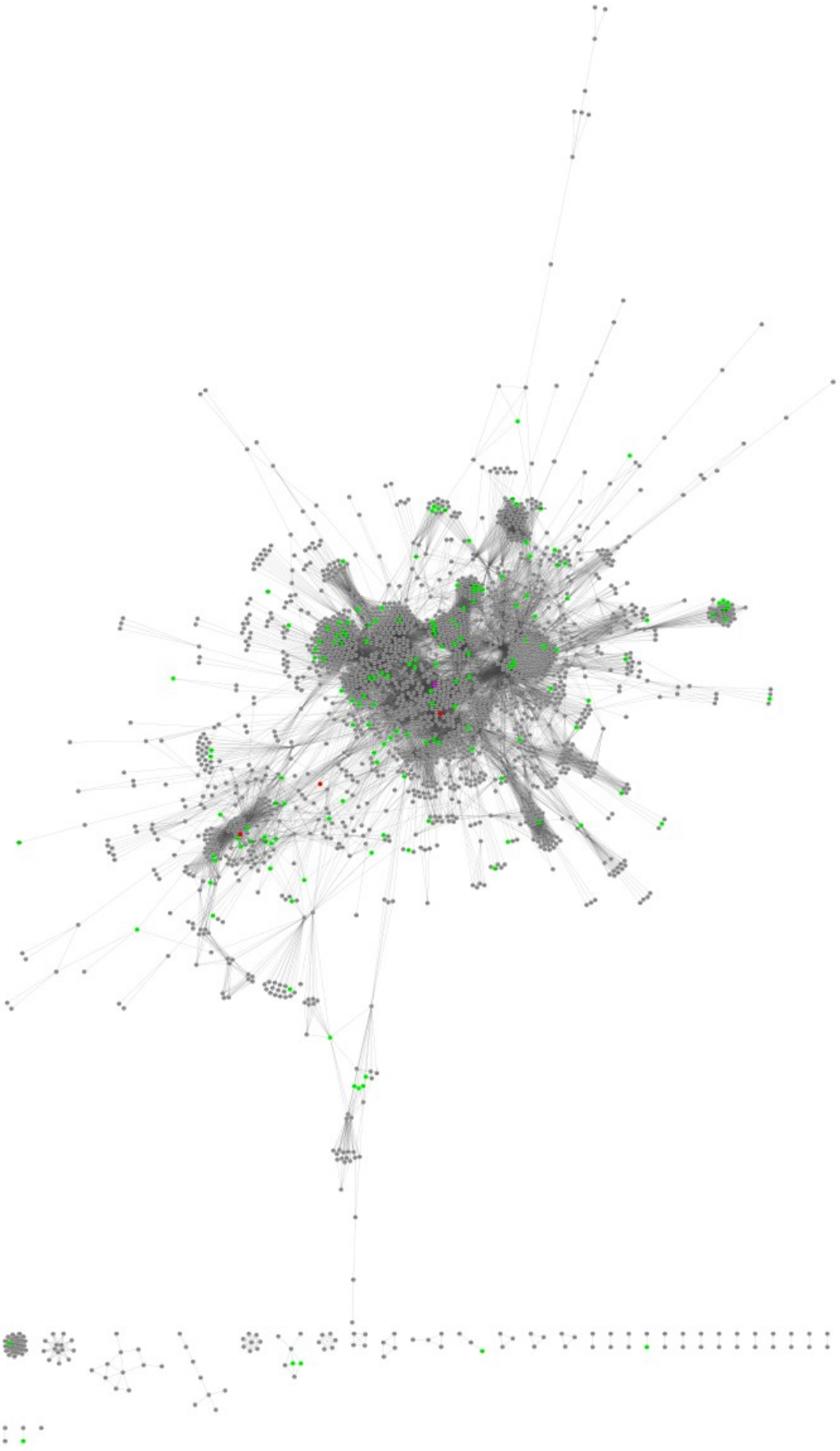


Fig. 19 - Distribution of genes affected by any kind of imbalanced variant. Green: genes specifically hit in the SR-positive patient group. Red: genes specifically hit in the SR-negative patient group. Violet: EGFR.

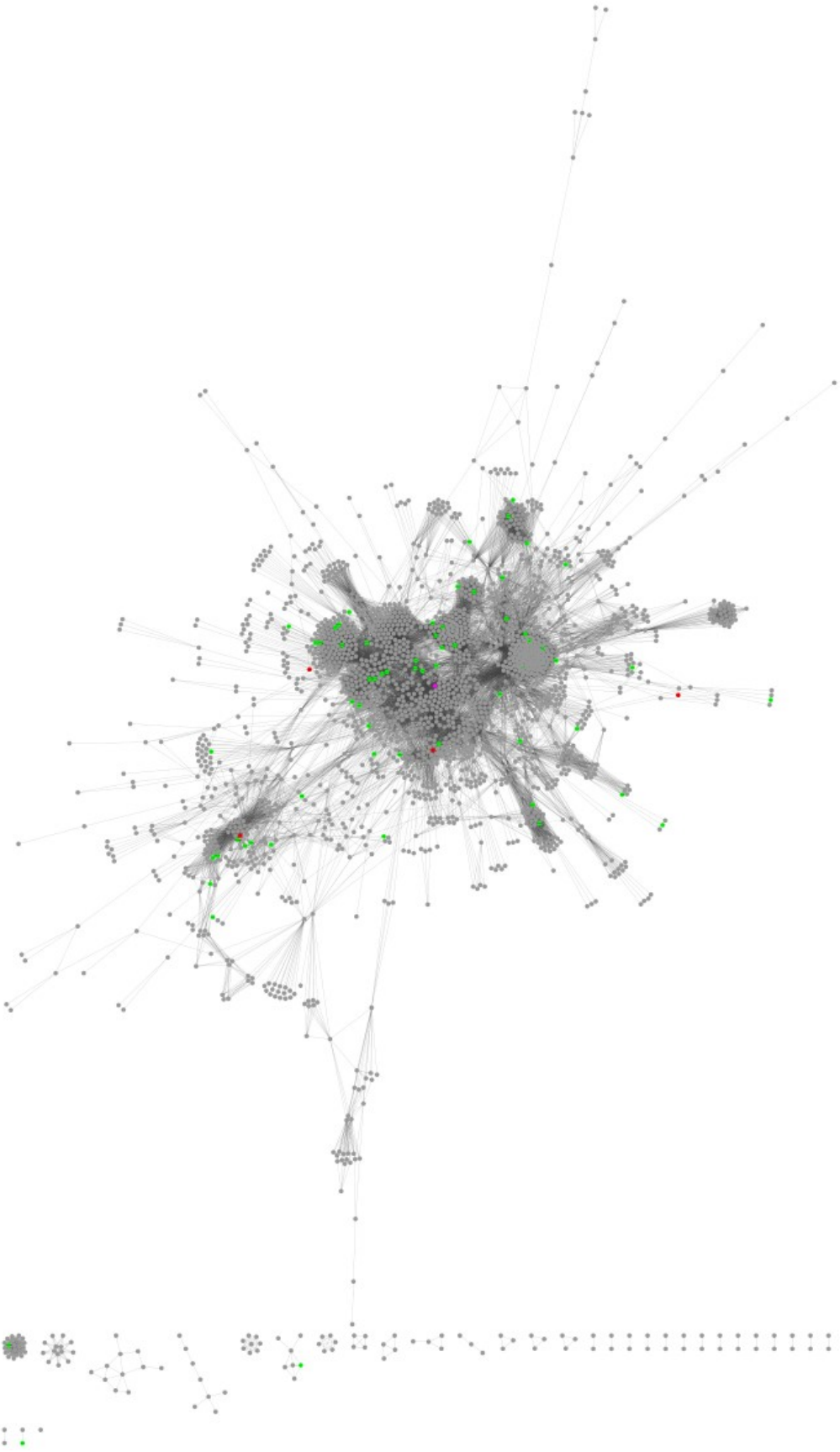


Fig. 20 - Distribution of genes affected by imbalanced MED or HIGH impact variant. Green: genes specifically hit in the SR-positive patient group. Red: genes specifically hit in the SR-negative patient group. Violet: EGFR.

When reducing the affected genes to those being hit by at least one medium (MED) or high (HIGH) impact imbalanced variant, it becomes feasible to compute the sets of shortest paths (SP clusters) to EGFR.

For Set 3 (Fig. 21), seven of those genes occurred. Six are assigned to the SR-positive patient group, with OBSCN (module 5) and STUB1 (module 3) being linked to EGFR with a distance of two functional interactions. CDH11 (module 7), COL4A4 (module 1), NUP210 (module 6) and TLR5 (module 4) show a distance of three. The single gene specific for SR-negative patients, P3H3 (module 1), has a distance of four and is directly linked to COL4A4. For those two collagen-associated genes, most pathes lead over PDGFA/B, NCAM1 or ITGB3 and few over ADAM10/17, connecting straight to the signaling-associated core of the MSM.

For Set 4 (Fig. 22), for five genes the SP cluster and shortest pathes to EGFR were computed. Again, P3H3 was hit and accounted for the SR-negative patient group. Here, NUP210 was assigned to this group, and not to the SR-positive group like in Set 3. The third gene, GRIP2 (module 2), shows a distance of two functional interactions to EGFR, with all shortest pathes leading via AP2 subunits (A1/2, B1, M1, S1). For the SR-positive group, two out of three specifically hit genes are related to clusters of differentiation: CD86 (module 0) is located in the MSM's center and closely connected to EGFR (distance = 2; same module). C3 (module 15) is connected to EGFR with several shortest pathes of distance three, and all these go via CD19. The third gene for this patient group is CCNK (module 10), connected exclusively via SMAD2/3/4 with SP distance 3.

3.5 Final candidate lists

When reviewing the variants detected by the pipeline, for some outputs minor allele frequencies (MAFs) appear to be greater than 0.5, effectively designating them to be the most common alleles for these loci. Conversely, the reference alleles are less common and by definition minor alleles. In contrast to the pipeline, dbSNP already lists the reference alleles' frequencies ($MAF_{reference}$) as official MAFs. For the most frequent case of two known alleles in a locus, this is simply $1 - MAF_{alternate}$. From a variant calling point of view, for these pipeline outputs the assignment of uncommon variants to a patient group has to be inverted: i.e., for C3's variant rs2287845 in Tab. 14, the less common 'G' allele with a MAF of 0.268 (= 1 - 0.732) has to be assigned to the SR-negative group.

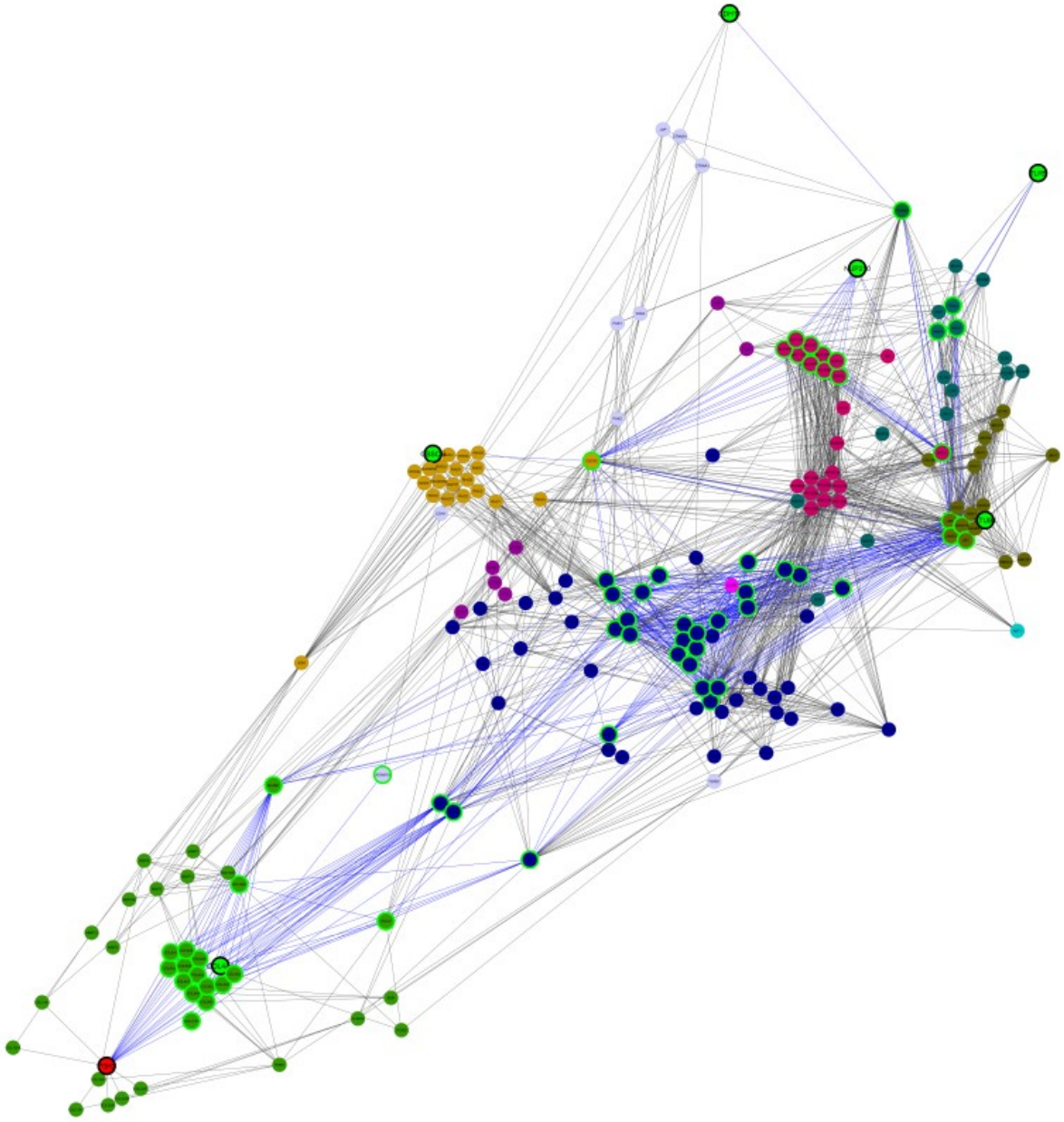


Fig. 21 - Clusters of Shortest Paths (SP) between affected genes of Set 3 and EGFR, each. Computed SPs are of a distance of two (OBSCN, STUB1), three (CDH11, COL4A4, NUP210, TLR5) and four (P3H3), respectively. Nodes on the SPs to EGFR are plotted with thick green borders, the connecting edges in blue. Further nodes linking all affected genes with each other via shortest pathes (= SP cluster) are painted in their respective module color (comp. Tab. 5), while respective edges stay grey. Green nodes with thick borders: with variants specifically observed in SR-positive cases. Red nodes with thick borders: with variants specifically observed in SR-negative cases. Violet node: EGFR.

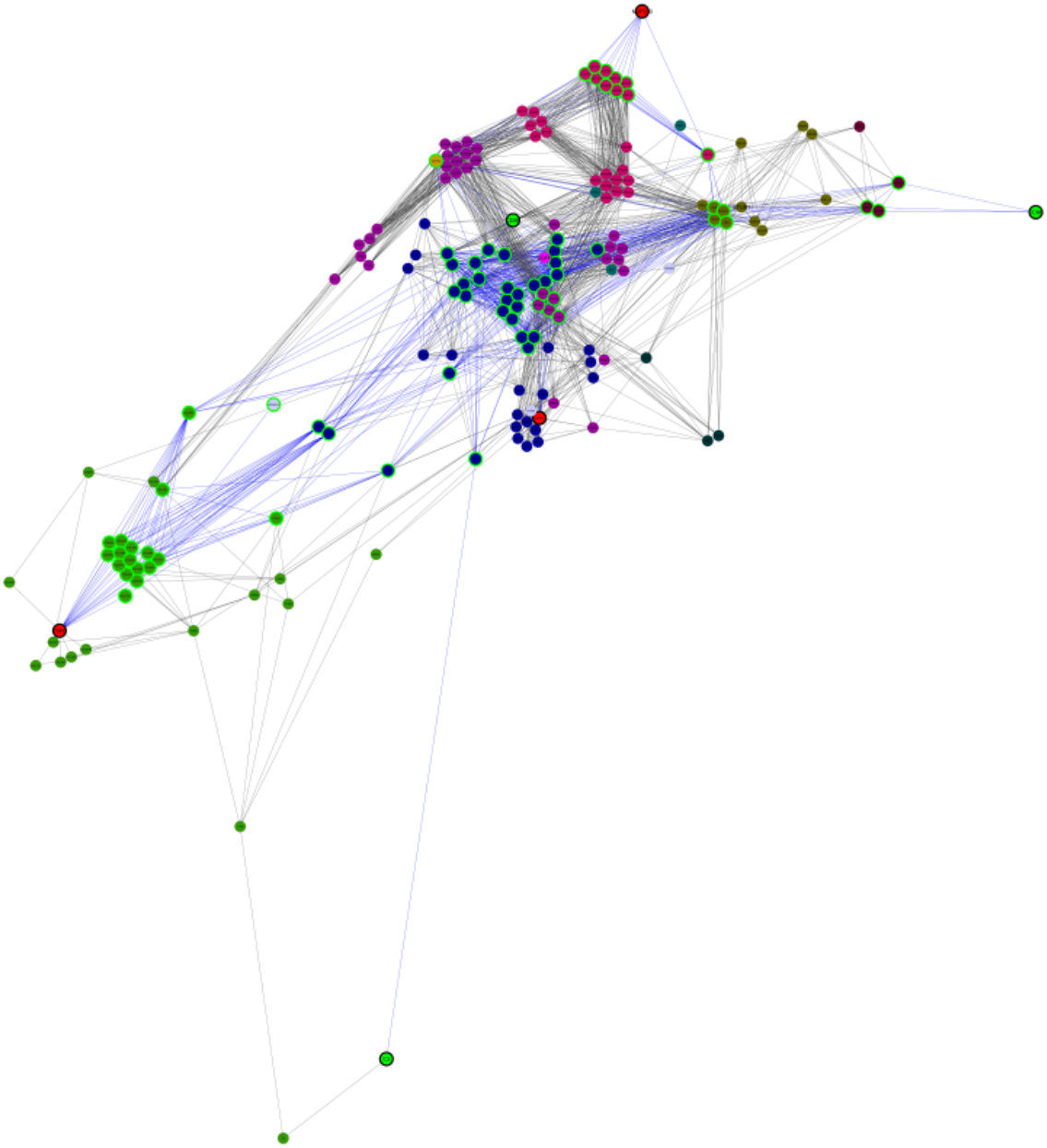


Fig. 22 - Clusters of Shortest Paths (SP) between affected genes of Set 4 and EGFR, each. Computed SPs are of a distance of two (CD86, GRIP2), three (C3, CCNK, NUP210) and four (P3H3), respectively. Nodes on the SPs to EGFR are plotted with thick green borders, the connecting edges in blue. Further nodes linking all affected genes with each other via shortest pathes (= SP cluster) are painted in their respective module color (comp. Tab. 5), while respective edges stay grey. Green nodes with thick borders: with variants specifically observed in SR-positive cases. Red nodes with thick borders: with variants specifically observed in SR-negative cases. Violet node in center: EGFR.

Tab. 12 - Reverse mapping from MSM to variants (Set 3). After restricting the set of genes affected by imbalanced medium or high impact variants of any type to those being part of the MSM, the following variants of interest remain. For pipeline outputs with minor allele frequencies (MAFs) greater than 0.5 (underlined), alternate alleles according to the reference have be considered as major alleles. The variant on P3H3 has been officially listed as rs57050687 recently and does not appear with this ID in the pipeline outputs. SO = Sequence Ontology.

Gene				Variant							Imbalance				
approved symbol [HGNC]	HGNC	Entrez	affected transcript count	genomic position	ref	alt	rs IDs [dbSNP]	MAF [aaf_1kg_all, dbSNP]	Impact [SO]	Severity [SO]	SR neg.	SR pos.	Assignment to SR group	Value	exclusive in group
CDH11	1750	1009	18	chr16:65025717	G	A	rs35195	0.182	missense_variant	MED	3	10	+	7	NO
COL4A4	2206	1286	2	chr2:227915831	G	A	rs1800517	0.547	missense_variant	MED	4	11	+	7	NO
NUP210	30052	23225	4	chr3:13361286	C	T	rs354478	0.612	missense_variant	MED	3	12	+	9	NO
OBSCN	15719	84033	15	chr1:228494789	G	A	rs435776	0.269	missense_variant	MED	3	10	+	7	NO
P3H3	19318	10536	13	chr12:6938022	C	CG	None (rs57050687)	1	frameshift_variant	HIGH	7	0	-	7	YES
STUB1	11427	10273	11	chr16:732284	AG	A	None	-1	splice_region_variant	MED	2	10	+	8	NO
				chr16:732286	GC	G	rs3216838, rs397766974	0.622	splice_region_variant	MED	2	10	+	8	NO
TLR5	11851	7100	6	chr1:223284527	A	G	rs386599677, rs5744174	0.29	missense_variant	MED	1	8	+	7	NO

Tab. 13 - Distribution of Set 3 variants across patients. For further information, refer to Tab. 12, using genomic position, reference (ref) and alternate (alt) alleles as key. "0" (green) codes for homozygous reference alleles, "1" (yellow) for one alternate allele detected, "3" (red) for homozygous alternate allele. "2" (grey), observed for chr16:732284 AG→A (in NUP210), codes for an unknown allele combination. Here, the genotypes are "AG/.", indicating for a mssing allele in the referring samples. All genotype codes follow the GEMINI convention.

Variant			Skin rash negative patients											Skin rash positive patients										
genomic position	ref	alt	020	036	072	125	281	375	406	428	586	750	796	014	090	137	155	111	213	344	566	598	624	638	708
chr16:65025717	**G**	**A**	0	3	0	0	0	0	0	1	1	0	0	1	0	1	0	3	1	1	1	3	1	3	1
chr2:227915831	**G**	**A**	0	0	0	1	0	0	0	1	1	3	0	1	1	1	3	1	3	1	1	1	1	0	1
chr3:13361286	**C**	**T**	0	3	0	0	0	3	0	0	1	0	0	3	1	3	3	3	3	1	1	3	1	1	1
chr1:228494789	**G**	**A**	3	1	0	0	0	0	0	3	0	0	0	1	3	1	3	1	1	1	0	1	0	1	1
chr12:6938022	**C**	**CG**	3	0	0	0	3	3	3	3	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0
chr16:732284	**AG**	**A**	0	0	0	0	0	0	0	0	0	2	1	1	2	0	1	2	1	2	0	1	1	1	1
chr16:732286	**GC**	**G**	0	0	0	0	0	0	0	0	0	3	1	1	3	0	1	3	1	3	0	1	1	1	1
chr1:223284527	**A**	**G**	0	0	0	0	0	3	0	0	0	0	0	1	1	1	1	1	0	0	1	1	0	0	1

Tab. 14 - Reverse mapping from MSM to variants (Set 4). After restricting the set of genes affected by imbalanced medium and high impact homozygous variants to those being part of the MSM, the following variants of interest remain. For pipeline outputs with minor allele frequencies (MAFs) greater than 0.5 (underlined), alternate alleles according to the reference have be considered as major alleles. The variant on P3H3 has been officially listed as rs57050687 recently and does not appear with this ID in the pipeline outputs. Entrez ID of GRIP2 (in italics) added manually, as not part of the offline database of SNPeff. SO = Sequence Ontology.

Gene				Variant							Imbalance				
approved symbol [HGNC]	HGNC ID	Entrez ID	affected transcript count	Position	ref	alt	rs IDs [dbSNP]	MAF [aaf_1kg_all]	Impact [SO]	Severity [SO]	SR neg.	SR pos.	Assignment to SR group	Value	exclusive in group
C3	1318	718	18	chr19:6696596	G	A	rs2287845, rs386490580	0.732	splice_region_variant	MED	3	10	+	7	NO
CCNK	1596	8812	9	chr14:99961829	T	C	rs2069492	0.447	splice_region_variant	MED	0	7	+	7	YES
CD86	1705	942	9	chr3:121825196	G	A	rs2681417	0.868	missense_variant	MED	4	12	+	8	NO
GRIP2	23841	*80852*	7	chr3:14551443	A	G	rs9845816	0.71	missense_variant	MED	7	0	-	7	YES
NUP210	30052	23225	4	chr3:13368891	G	A	rs2271509	0.396	synonymous_variant	LOW	7	0	-	7	YES
				chr3:13383539	A	G	rs2271504	0.577	synonymous_variant	LOW	7	0	-	7	YES
				chr3:13395578	C	A	rs2280084	0.575	missense_variant	MED	7	0	-	7	YES
P3H3	19318	10536	13	chr12:6938022	C	CG	None (rs57050687)	1	frameshift_variant	HIGH	7	0	-	7	YES

Tab. 15 - Distribution of Set 4 variants across patients. For further information, refer to Tab. 14, using genomic position, ref and alt alleles as key. “0” (green) codes for homozygous reference allele and “3” (red) for homozygous alternate allele. All genotype codes follow the GEMINI convention (“0” = homozygous reference; “3” = homozygous alternate).

Variant			Skin rash negative patients											Skin rash positive patients											
Position	**ref**	**alt**	**020**	**036**	**072**	**125**	**281**	**375**	**406**	**428**	**586**	**750**	**796**	**014**	**090**	**111**	**137**	**155**	**213**	**344**	**566**	**598**	**624**	**638**	**708**
chr19:6696596	**G**	**A**	3	0	0	0	0	0	3	0	0	0	3	3	3	3	3	3	3	3	3	0	3	3	0
chr14:99961829	**T**	**C**	0	0	0	0	0	0	0	0	0	0	0	3	0	0	3	0	3	3	0	0	3	3	3
chr3:121825196	**G**	**A**	0	3	0	3	0	3	3	0	0	0	0	3	3	3	3	3	3	3	3	3	3	3	3
chr3:14551443	**A**	**G**	3	0	0	0	3	3	0	3	3	3	3	0	0	0	0	0	0	0	0	0	0	0	0
chr3:13368891	**G**	**A**	3	0	0	3	3	0	3	3	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0
chr3:13383539	**A**	**G**	3	0	0	3	3	0	3	3	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0
chr3:13395578	**C**	**A**	3	0	0	3	3	0	3	3	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0
chr12:6938022	**C**	**CG**	3	0	0	0	3	3	3	3	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0

4. Discussion

Skin rash as an undesired, nowadays often preventively treated side effect, has been used as a phenotypic biomarker for drug efficacy of Cetuximab (and other EGFRIs) by oncologists for years now. A minority of patients, who more or less lack a skin toxicity reaction, are reported to be low responders to Cetuximab. Although this repeatitively reported correlation is still under investigation and not finally confirmed, in practice SR-negative patients are subject to switch of therapy.

However, molecular mechanisms potentially connecting these two observations causally, are still unknown. Apart from scientific motivation, deeper knowledge may support optimized treatment strategy selection by earlier decision for alternate medications.

As hypothesis generation is an important first step, exome data from eleven SR-negative CRC patients and twelve SR-positive controls, treated in first line with Cetuximab each, were used as the basis for this work. After state-of-the-art variant calling from raw NGS data, two approaches were combined for finding candidate genes. First, following an assumption that relevant affected genes should be significantly, inequally distributed between both patient groups, the 'imbalance' was set up as a statistical model. Second, by restricting the subsequent analyses to a set of biologically promising mechanisms, candidate genes should be filtered massively. Therefore, a usecase-specific 'Molecular Systems Map' (MSM) was created, while the MSM principle itself is generic. The MSM has been implemented in Cytoscape, which is most commonly used in bioinformatics. Here a focus is on molecular and genomic interactions in curated, uncurated as well as custom made networks, which can be visualized and analysed simultaneously in a graph-oriented and semantic-aware manner [Fitts *et al.* 2016].

4.1 Variant calling and the imbalance criterion

In previous analyses, performed by the NGS data-producing lab, no simple relation between a single gene and the observed phenotype could be found. Consequently, for this work two central assumptions were made:

- Not a single gene is responsible for the phenotype across all affected patients, but maybe a functionally connected subset.
- Exome data in fact is enriched for exons by design, but the given short reads may be generated from further loci or exceed the target regions, carrying additional information.

Therefore, no BED file was used for the short read mapping, leading to a larger number of called variants in intronic or even intergenic regions. Here, functional elements like promoters, various binding motifs and splice sites are expected to occur, having influences when changed by somatic variations. Calling technical artefacts is considered to be compensated by the removal of read duplicates.

4.2 Evaluating the skin rash MSM

The MSM provides a visual, interactive representation of those mechanisms and a topologic representation of the detected genes, e.g. their direct interaction partners, connection to EGFR and membership to functional units (pathways, clusters of functional interactions). Apart from gaining deeper insights in the biological meaning of the NGS-derived results, hypothesis generation by domain experts (physicians, geneticists) will be supported.

For the given use case of skin rash in Cetuximab treatment, the MSM fulfilled the attributes stated above and delivered reasonable structures when considering the data-driven layout and clustering compared to the biological backgrounds. In the layout (for a detail shot depicting the following paragraphs refer to Fig. 23), **dense clusters** represented tightly connected functional units, which are simultaneously accounted for respective modules by ReactomeFI's algorithm. Much more important, modules can be assigned easily for biological functions. For sure, on the one hand, as Reactome's pathway annotations are manually curated and follow the biological knowledge, these observations meet an expectation. But, on the other hand, the generated MSM layout follows this principle expectations, although Reactome data is most often used in a pathway-centered way.

Furthermore, the **central, but less tightly clustered modules** are distributed in the core region (module 0; growth factor signaling and cytokines) or even across major parts of the MSM (module 7; cell interaction/junction upon Rho/TGFβ), showing their connective character. This makes biologically sense, as growth factors share multiple components and perform intensive cross-talk [Holland *et al.* 2003], acting more as a reaction network – in turn underlining the artificial character of the 'pathway' concept.

Generally, the **top left quarter of the core region** shows the close interaction of components from both immune system and growth factor/GPCR-related signaling pathways, all primarely related to sensing conditions from the cell's environment (extracellular space). In terms of immunology components, Fcγ Receptor-mediated phagocytosis (module 12; innate immunity) can be found here, while the major parts are contributed by the mechanisms of cytokine signaling, especially interleukin-

like ones (belonging to module 0). Those are released into the intracellular space during inflammation reactions and sensed by receptive cells. Additionally, MHC class II signaling (module 2) as the respective component in the adaptive branch of the immune system, adjoins here. These molecules are present primarely in professional antigen-presenting cells, like dendritic and B cells, providing antigens from phagocytosed proteins. In contrast, the MHC I cluster is located much more right within the MSM, showing less proximity to the core region.

Close to module 0, but quite strictly separated, components of **nucleus mechanics** (module 6; nuclear pores, chromatids) appear, reflecting that downstream events of both growth factor signaling and immune system reactions include highly regulated transcription factor re-localization from cytoplasm to nucleus.

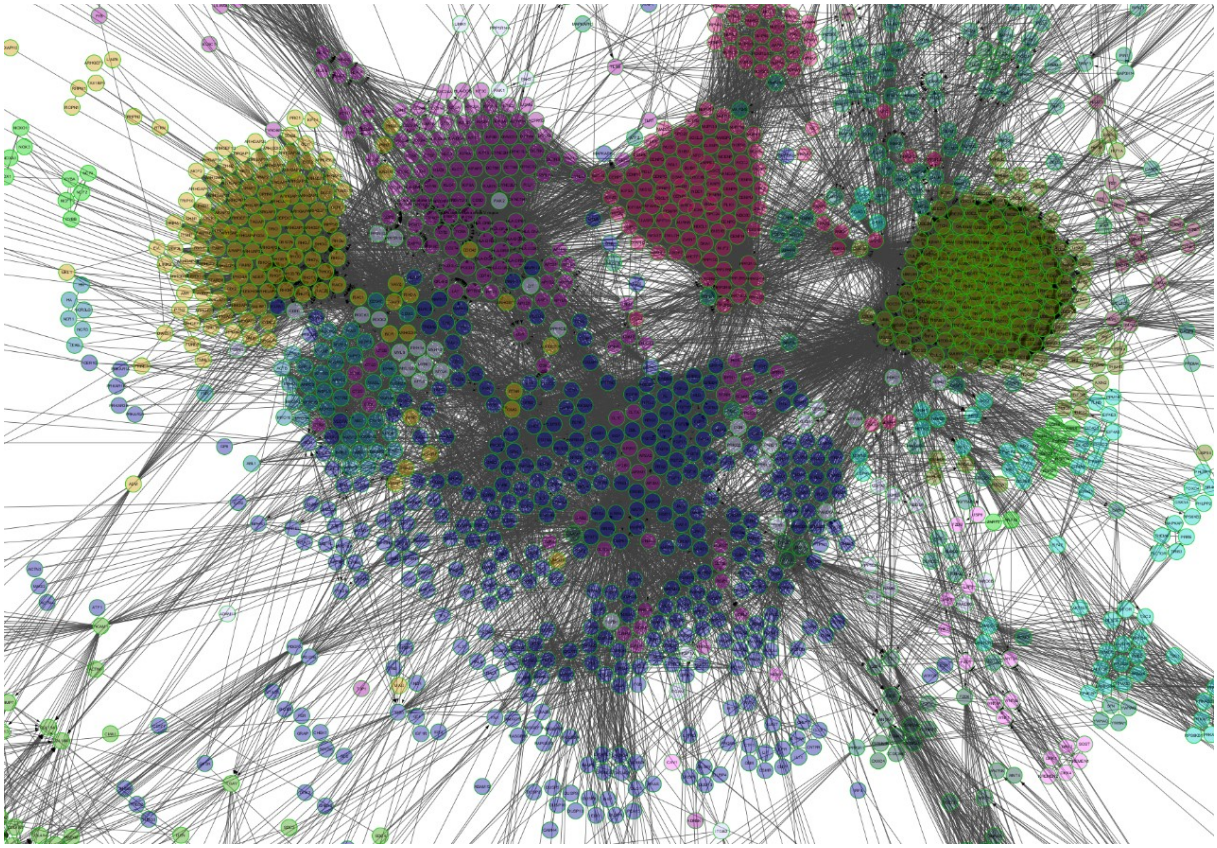


Fig. 23 - Detail from the MSM's central region. Example for the relation between modules 0 (blue), 2 (violet), 3 (olive), 4 (light blue), 5 (yellow), 6 (pink/magenta) and 7 (violet grey).

However, there were some surprises when interpreting the MSM, e.g. when checking for the localization of cell adhesion-related molecules. Cadherins appear topologically separated from all other components from this superset (which form module 1 at the bottom left side of the MSM), located closer to WNT-related modules on the top right side. Biologically, this may underline the tight

mechanistic connection to WNT pathways (especially in development), as well as the dual function of cadherins as mechanical connectors and receptors/ligands for signaling.

Also, when considering the WNT-related reactions, a separation into three major modules (8, 11, 13) was not expected. While module 13 may map properly to the WNT ligand biogenesis und trafficking pathway, the modules 8 and 11 share larger intersections. In more detail, module 8 appears like integrating all WNT-related functions including the downstream link to nuclear factors (transcription factors, β-Catenin, Demethylases, histones; also assigned to Signaling by Rho GTPases pathway). Module 11's connection to the RAF/MAP kinase pathway remains functionally unclear (G-protein receptors, voltage-gated potassium channels).

Most interestingly, EGFR as Cetuximab target molecule and consequently the starting point of all literature research on connected molecular mechanisms, appeared right in the middle of the MSM. Obviously, as it belongs to the central growth factor signaling representing module 0, it was not expected to be located in any corner of the map. But from a functional point of view, for the given subset it takes a literally central role, as in contrast to e.g. receptors for foreign molecules (TLRs, LRRs), there it not only a downstream side from the receptor, but also a number of regulative mechanisms on the ligand side, pointing to EGFR. Considering that the receptor knows a number of related but different endogenous ligands (EGF, HB-EGF, … [Wells 1999]), it is not a simple starting point of a cellular reaction, but central switch of multiple regulative cycles, obviously pulling the respective node into the MSM core.

Apart from biological interpretation, there is a technical side on the MSM's implementation in Cytoscape. During development some technical limitations were experienced, mainly due to the amount of data – although highly connected nodes (e.g. rhodopsins, immune chains) were removed beforehand during pathway merging. In terms of performance, especially the heavy usage of main memory is grave: while up to eight gigabytes are used by Cytoscape in the given setting, usual desktop computers with less than 16 GB of RAM might struggle. Here, the user's experience might be rough. Ultimately, for this reason, excessively connected subsets of nodes were removed from the initially loaded pathways (comp. Subsection 2.7.1 and Tab. 4), reducing objects (nodes and edges) to a number operable by Cytoscape.

Furthermore, by observation somehow connected to the memory consumption, sessions in Cytoscape might loose contents (e.g. plotting styles, edge annotations fetched from ReactomeFI) or functionalitites (e.g. right-click menu in the main panel, with all its functions). Also freezes of the whole workbench were faced, and saving sessions takes several seconds to finish for sessions containing many objects.

Independent of those issues, the statement has to be made that in terms of usability and non-expert user motivation, Cytoscape as an interface is still a *de facto* standard for graph data.

In summary, the MSM idea of integrating functional interaction data works, at the model follows the annotated biological subject, designating the MSM as a valid basis for the analyses which incorporate the NGS-derived variation data.

4.3 Final gene candidates and interpretation

When reviewing the results of the performed analyses on the exome data from a systemic perspective, for Sets 1 and 3, representing those which still include the majority of variants being designated by SNPeff as low impact, no eye-catching distribution pattern could be observed.

In contrast, the restriction to medium or high impact variants in both Set 2 and 4 revealed a small number of remaining genes, suitable for manual review within the MSM itself as well as further knowledge bases like GeneCards[13] [Rappaport *et al.* 2017], dbSNP [Kitts *et al.* 2013][14], PubMed[15] and others. For those candidates, the following subsections introduce findings and, where applicable, basic hypotheses on how the variant may explain the Cetuximab-related skin rash phenomenon.

Taken together, the Sets 2 and 4 provide eleven candidate genes, of which ten are discussed the following Subsections 4.3.1 to 4.3.10. With KISS1, extracted from the NGS data variation analysis, but missed by the MSM design, another gene will be discussed in Subsection 4.3.11.

4.3.1 C3

C3 plays a central role in activating the complement system, which works in balanced cascades of proteolytic steps and potentiates antimicrobial responses of both humoral and cellular nature [Merle *et al.* 2015a; Merle *et al.* 2015b]. Mechanistically, three sensoring pathways (classical, alternative and lectin pathway) join in C3's amplification loop, finally initiating the terminal pathway and its membrane attack complex [Merle *et al.* 2015a]. In recent years the complement system's role has been recognized as being more than microbial defense [Mastellos *et al.* 2016]. It plays a central role for tissue homeostasis, also via the recognition of diseased and damaged cells of the host itself, together with other immune mechanisms [Ricklin *et al.* 2010]. An important cleavage product is C3b, acting as opsonizing factor, tagging cells for decay and recruiting phagocytotic cells; C3a meanwhile acts as

[13] http://www.genecards.org
[14] https://www.ncbi.nlm.nih.gov/SNP/
[15] https://www.ncbi.nlm.nih.gov/pubmed/

anaphylatoxin, triggering potent inflammatory responses in the local tissue [Janeway 2001]. Interestingly, in cancer settings, C3 is described to be elevated in the tumor microenvironment, and having an anti-angiogenic effect [Pio *et al.* 2014]. Meanwhile, lowered C3b levels seem to be correlated with an improved escape of the tumor from immunosurveillance [Pio *et al.* 2014]. In human, but not mouse skin, keratinocytes treated with EGFRIs showed elevated expression and activation of complement components, together with improved recruitment of neutrophils, implicating a chemotactic effect [Holcmann & Sibilia 2015; Abu-Humaidan *et al.* 2014], leading to inflammation and rash [Holcmann & Sibilia 2015]. Since more than thirty years, the relation between C3 polymorphisms and Crohn's disease, a chronic inflammatory bowel disease (IBD) is documented [Elmgreen *et al.* 1984].

In the given SR setting, C3 occurred to be hit homozygously by a variant altering a splice site in the SR-positive patient group. As the minor allele frequency (MAF) was given with 0.73, in fact, the reference genome is carrying the less frequent variant, finally switching relations: the SR-negative patient group carries in eight out of eleven cases the less frequent allele of a variation site. In contrast to the dbSNP information used in the data processing/annotation, this variant (rs2287845) has already been switched to the described opposite case in dbSNP online, with a MAF of 0.27. Furthermore, it is tagged with a clinical relevance, pointing to C3 deficiency (NCBI MedGen ID C1332655). The ClinVar entry[16] states 'benign' for the more common allele (which is present in the SR-positive group). In this context it is announced that patients with a C3 deficiency may show systematic lupus erythematosus (SLE; [Tsukamoto *et al.* 2005]). SLE, also inducible by drugs, shows a similar phenotype to the observed Cetuximab-induced skin rash [Antonov *et al.* 2004].

Interestingly, Chowdhury *et al.* 2015 already reported a C3 polymorphism to have a significant influence on a treatment. Despite the fact that this publication is about a coding variant and hepatitis C, it hits quite fair the pharmacogenomics principle, which is part of the working assumptions of this thesis. In another study referenced at MedGen, there are three coding variants reported affecting the exons 24 (frameshift) and 26 (premature stop codon, nonsense mutation), respectively [Kida *et al.* 2008]. An additional study reported about a splice site mutation, leading to a skip of exon 27, compromising a functional domain, and ultimately leading to a complete C3 deficiency [da Silva *et al.* 2016]. In turn, the patient reported in [Tsukamoto *et al.* 2005] was identified with an acceptor site A-to-G transition in intron 38, which leads to skipping exon 39. The parents and two sibblings were heterozygous for this mutation and showed reduced levels of C3 hemolytic activity. Taken together, a number of clinically relevant C3 variants are known.

[16] https://www.ncbi.nlm.nih.gov/clinvar/variation/330297/

For the variant rs2287845 detected in the given skin rash setting, neither a clinical phenotype nor an elevated interest has been reported yet. Here, the affected donor splice site reported by SNPeff is located after exon 22. Assuming the splice out signal is impaired, the intron region up to exon 23 would remain part of the mature mRNA, getting finally translated into an amino acid sequence. As the additional region of 127 bp is not divisible by three, the subsequent nucleotide sequence would be affected by a frame shift event. As shown in Box 12, 14 deviating amino acids would be coded until a premature stop codon terminates the translation.

Box 12 - Predicted sequence change in C3 due to rs2287845. The original sequence of the mature mRNA contains the transcribed information from exons 22 (yellow highlight) and 23 (cyan highlight), being translated into the reference amino acid sequence (as shown in bold green). With the reference "c" in intron 22 (lower case subsequence; highlight in grey) converted to a "t" (highlighted in red), the resulting mutation event "NG_009557.1:g.29066C>T" causes a disruption of the donor splice site as predicted by SNPeff. Thus, the intron is probably not spliced out.
Consequently, the amino acid sequence is also changed. In the reference sequence of the mature mRNA[17], the codon "GAA" starting from position 2957 (underlined) is assembled of one base from exon 22 and two bases from exon 23. It codes for glutamic acid ("E") at position 956 of the reference protein sequence[18]. When translating the not spliced out intron, the amino acid sequence would start deviating with a frameshift from the reference (red bold), introducing a mutation p.E955G (with "G" coded by new codon "Ggt"; underlined). The 14th following triplett would finally code for a new stop signal ('*'), terminating the translation here. Taken together, the modern HGNC nomenclature would code the event on the protein level as "NP_000055.2:p.(Glu955GlyfsTer969)". Note: DNA sequence is the reverse complement of the reference, as C3 is coded on the reverse strand.

original mRNA sequence

```
2890 CCGGAAGGAATCAGAATGAACAAAACTGTGGCTGTTCGCACCCTGGATCCAGAACGC 2947
 933   P   E   G   I   R   M   N   K   T   V   A   V   R   T   L   D   P   E   R   952
2948 CTGGGCCGTGAAGGAGTGCAGAAAGAGGACATCCCACCTGCAGACCTCAGTGAC    3000
 953   L   G   R   E   G   V   Q   K   E   D   I   P   P   A   D   L   S   D       969
```

modified sequence after splice site disruption

```
2890 CCGGAAGGAATCAGAATGAACAAAACTGTGGCTGTTCGCACCCTGGATCCAGAACGC 2947
 933   P   E   G   I   R   M   N   K   T   V   A   V   R   T   L   D   P   E   R   952
2948 CTGGGCCGTGgtgagttggctgcagggggaggggctgagggctggcagggtaaggg    ?
 953   L   G   R   G   E   L   A   A   G   G   G   A   E   G   L   A   G   *   G     ?
   ?  gggtaaatgacctggttagtgaggtaggataggggcgggagggagctagagccatc    ?
   ?   G   *   M   T   W   V   *   *   G   R   I   G   R   E   G   A   R   A   I     ?
   ?  ggtatctctcactcaccctgcagAAGGAGTGCAGAAAGAGGACATCCCACCTGCAGA    ?
   ?   G   I   S   H   S   P   C   R   R   S   A   E   R   G   H   P   T   C   R     ?
   ?  CCTCAGTGAC
   ?   P   Q   *
```

Considering the position of the variation at ~60% of the 1,663 bp coding sequence of C3, a truncated version of the protein as considered in Box 12 may be regarded as too strong. The SR-negative patients show the variant homozygously, and are not described as phenotypically different to other patients before treatment. C3 is a central player in all three complement cascade pathways and the SLE-like phenotype of complete C3 deficiency as announced by [Tsukamoto *et al.* 2005] has not been documented for the given CRC patients. Consequently, it would be reasonable to assume a new donor

[17] https://www.ncbi.nlm.nih.gov/nuccore/NM_000064
[18] https://www.ncbi.nlm.nih.gov/protein/NP_000055

splice site appearing in the unspliced intron region before the newly introduced stop codon. The resulting protein would accordingly not lack major parts, but show a change in its threedimensional conformation, as additional amino acids would occur compared to the wildtype form. From a functional domain point of view, NCBI's Conserved Domain Database search engine[19] (CDD; [Marchler-Bauer *et al.* 2017]) reports a 'complement_C3_C4_C5' domain[20] starting just 31 amino acids after the affected position (Fig. 24). The respective domain is central for C3, as it contains four important features (active site, thioester region, surface patch and specificity-defining residues) central for binding of and reaction with other proteins. Presumably, while still present and intact, the domain might be suboptimally oriented within C3's three-dimensional structure, affecting e.g. the binding to factors B or H, the cleavage or the probability to exhibit the thioester site during 'tick-overs' (see down below).

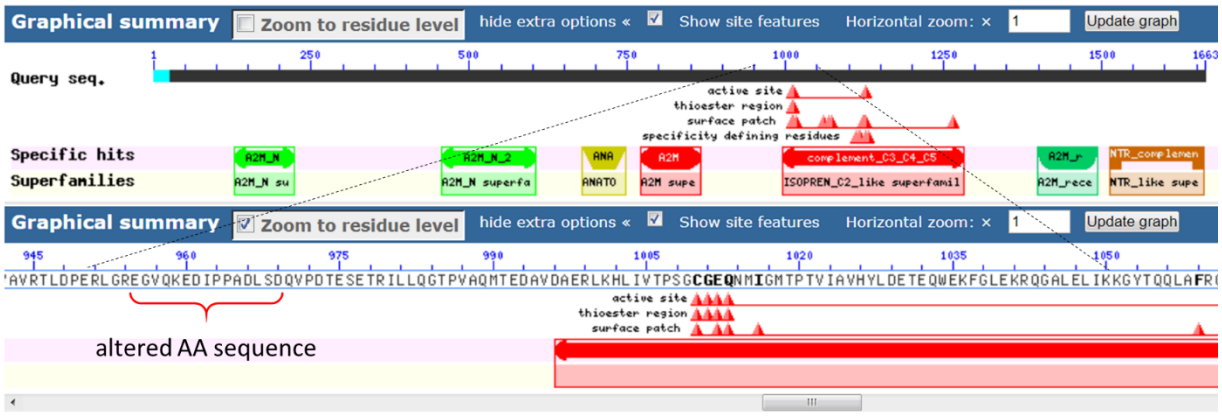


Fig. 24 - Conserved Domain Database (CDD) view on C3's protein sequence. The broader red box marks the 'complement_C3_C4_C5' domain, which follows shortly after the altered region in which an unspliced intron potentially introduces additional amino acids an a frameshift (red bracket). Top half: full domain-centric view onto protein sequence; bottom half: zoom to region of interest. Amino acids printed in bold carry dedicated functions in 3D space.

In a nutshell, for the SR-negative patients showing the rs2287845 alternate allele, C3 is probably somatically slightly impaired, but not completely dysfunctional. On the one hand, this would imply changes in the regulative potential due to altered conformation may trigger large-scale changes in immune reactions overall: C3 is modulatory active on both adaptive and innate immune responses, affecting human pathophysiology via diverse biological processes [Mastellos *et al.* 2016]. On the other hand, a highly regulated and usually robust biological system like immunity characterizes itself by deploying adequate mechanisms [Azeloglu & Iyengar 2015], constantly balancing between minimizing damage by pathogens and autoimmunity [Bergstrom & Antia 2006; Eberl 2016]. Feedback control and

[19] https://www.ncbi.nlm.nih.gov/Structure/cdd/docs/cdd_search.html
[20] https://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=cd02896

redundancy are just two well-known examples described in several signaling cascades and pathogen elimination strategies [Bergstrom & Antia 2006]. In fact, for the complement system a clinically relevant long-term compensation for primary deficiencies in C3 has been reported: while those individuals show an increased risk for infections at young ages, in adulthood this effect subsides, suggesting a yet unknown mechanism [Reis *et al.* 2006]. Thus, investigations on the medical history of SR-negative patients may be informative.

For a mechanistic overall hypothesis regarding C3's role in the SR setting, it should be noticed that EGFR inhibition leads to an elevation of C3 levels in keratinocytes [Abu-Humaidan *et al.* 2014]. In the skin, which is in a regular contact to the microbial environment, inflammatory cascades are triggered constantly via the alternative pathway. Simultaneously, the activating forces are counteracted by inactivation processes, keeping the whole system in a delicate balance [Zipfel 2001]. Usually, this dynamic equillibrium resides on a low level, providing homeostasis under physiological conditions [Harris *et al.* 2012]. When due to Cetuximab's induction, C3 molecules are more abundant than in a non-treatment setting, its higher concentrations may lead to more molecules being initially activated (Fig. 25). As consequently more units of the cascades subsequent steps are present, the probability for the whole complement cascade to 'fall over' would increase, effectively pushing the homeostasis to inflammation [Harris *et al.* 2012]. This would finally trigger and amplify immune cell recruiment and cytokine actions, resulting in skin rash [Holcmann & Sibilia 2015].

In consequence of C3's role in homeostasis, pathogenesis of a wide spectrum of inflammatory diseases is closely related to dysregulation of the complement system, mostly described as overactivation and including some individuals showing a skin rash phenotye similar to SLE [Ricklin & Lambris 2013; Truedsson *et al.* 2007]. Regarding therapeutic intervention, silencing of hypersensitive complement cascade reactions by blocking C3 as central component, is already subject to ongoing studies [Ricklin & Lambris 2016]. In the given CRC treatment setting, for C3 wildtype patients this overreaction may be triggered exogenously by systemically applied Cetuximab, elevating the levels of cleavable C3. As a result, the immune equilibrium of the skin gets massively disturbed, leading to a pathogenic state [Eberl 2016].

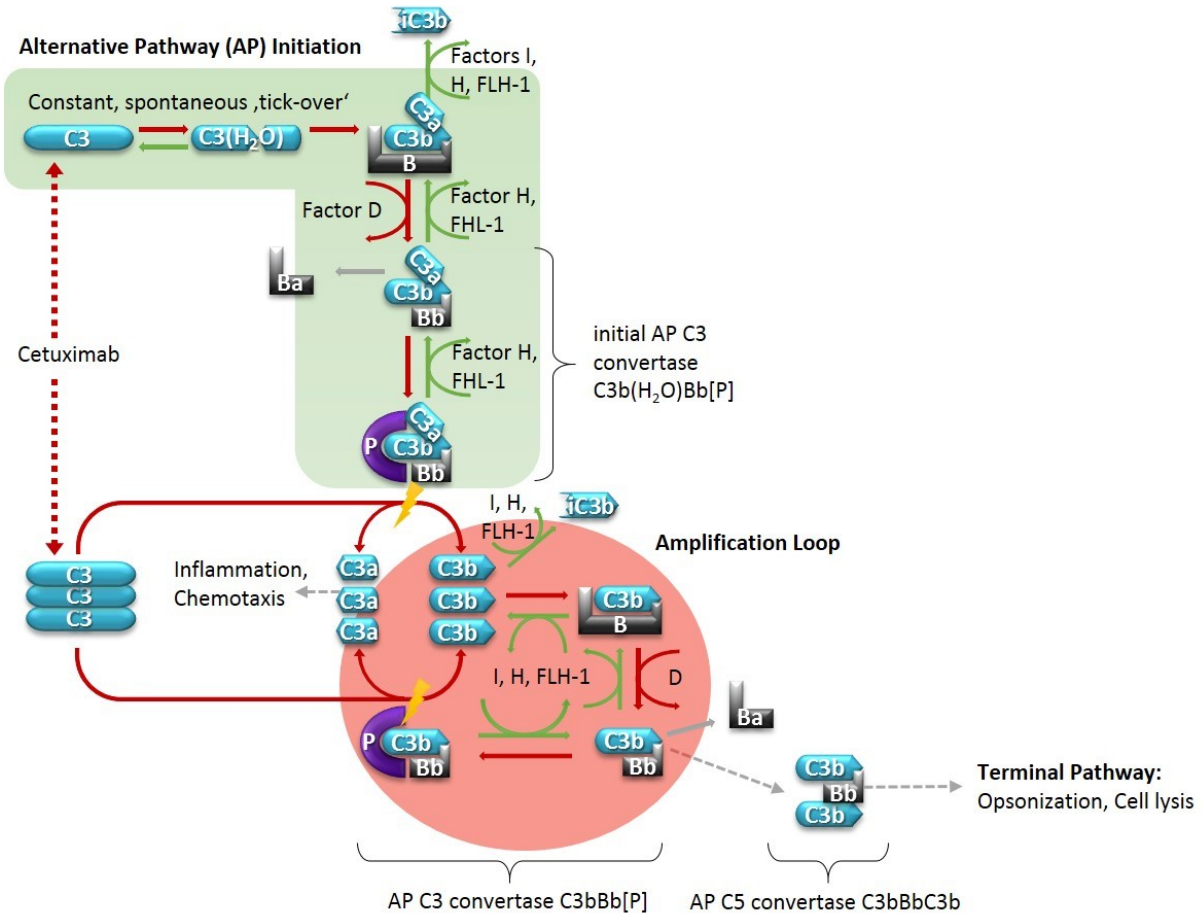


Fig. 25 - Simplified model of C3's hypothetical role in skin rash phenomenon via the alternative pathway (AP). In the AP, C3 molecules (blue) spontaneously convert to bioactive C3(H_2O). This hydrolysis is known as 'tick-over' [Merle *et al.* 2015a]) and is the starting point for the AP initiation (green area). Hydrolized C3 is structurally different, effectively exposing a binding site for Factor B (black). Bound Factor B in turn gets cleaved by Factor D, leaving C3(H_2O)Bb, which is capable of binding stabilizing Factor P (violet). The initial C3 convertase of the alternative pathway (AP) irreversibly cleaves a number of C3 molecules, which again interact with Factor B, making up an amplification loop (red area; [Lachmann 2009]). During activation, most steps forward (red solid arrows) are counteracted by Factor H and FHL-1, representing a competing inactivation process (green solid arrows), effectively creating a dynamic equillibrium between activation and inactivation. For the later, the non-reversible exit point for C3b is the cleavage by Factor I, resulting in iC3b (both in initiation and amplification). In contrast, upon sufficiently high turnover rates, C3 convertases in the amplification loop bind free C3b, resulting in C3bBbC3b complexes (the C5 convertase of the AP). Those finally trigger the immunologically effective terminal pathway.
When considering Cetuximab's effect to increase C3 expression, both initiation and amplification of the AP could be assumed to be promoted, as turnover rates get increased and the equillibrium gets influenced in favor of the system's forward direction. A general shift to activation would be the consequence for SR-positive patients, leading to inflammation (rash). Considering a slightly impaired C3 variant in the SR-negative patients, protein interactions on all steps of the cascade could be affected, presumably causing a reduced reaction speed in the progression to inflammation (interaction with Factors B, D or P). Consequently, the equillibrium gets shifted backward to inactivation and no inflammation (no rash). Illustration and general mechanistic description following [Zipfel 2001] and G. Hegasy[21].

[21] http://www.hegasy.de (Downloads > Komplement System)

For the tumor, the complement systems comes into play with contrasting positions; acute inflammation, lysis, chemotaxis and other mechanisms control tumor activities, while chronic inflammation promotes immunosuppression in the tumor's microenvironment, angiogenesis and progression [Khan *et al.* 2015]. Especially for the colon, connections have been drawn between chronic inflammation (inflammatory bowel disease; IBD) and an increased risk of cancer (colitis-associated cancer; CAC). Here, the balance shift from physiological (acute) to pathological inflammation (chronic) is ground-laying for a tumor-supportive microenvironment [Danese *et al.* 2011]. Considering the colon, and especially the colonic tumor [Pabla *et al.* 2015], as EGFR-expressive tissues [Wells 1999; Cohen 2003], elevated C3 levels in response to EGFR shutdown [Abu-Humaidan *et al.* 2014] would lead again to a homeostasis shift of the complement system's alternative pathway in favor of acute inflammation, like for the skin's keratinocytes. In the first line, a result would be a more sensitive recognition of malignant cells, supporting the anti-tumor forces of the immune system [Pio *et al.* 2014]. A report on Cetuximab's activity to be significantly higher when the de-activating Factor H was genetically downregulated, supports this model (study in lung cancer; [Hsu *et al.* 2010]). Moreover, the same authors reported a general anti-tumor activity of Cetuximab being linked to complement activation via the classical pathway, resulting in increased production of cytolytic C5b-C9 complex (endpoint of the complement system's terminal pathway). Consequently, apart from the alternative pathway's influence described above, especially for the tumor context the complement's classical pathway has to be considered relevant for the skin rash phenomenon (Fig. 26).

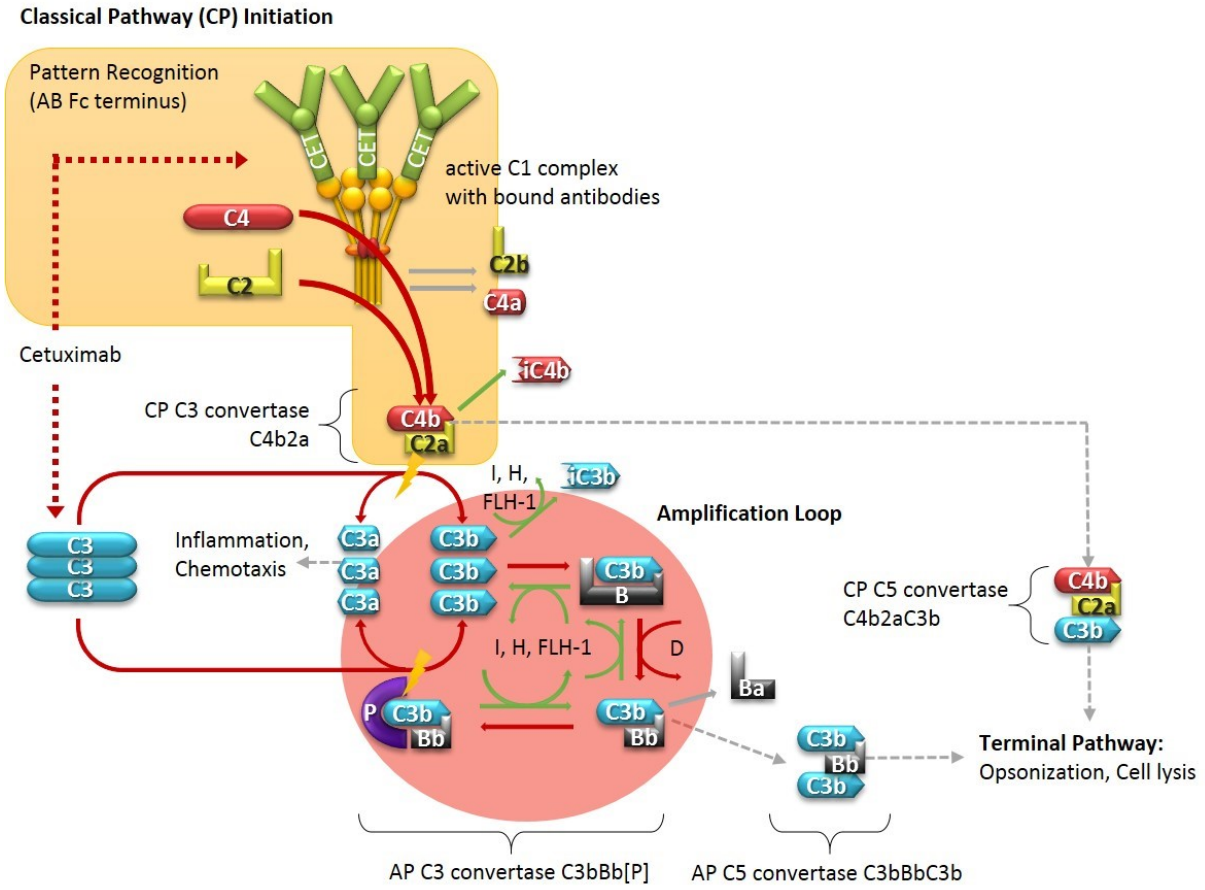


Fig. 26 - Simplified model of C3's hypothetical role in skin rash phenomenon via the classical pathway (CP). In the CP, no spontaneous activation occurs: initiation (orange area) is triggered by the recognition of foreign patterns by antibodies of the IgM or IgG type. The antibodies Fc terminus is subsequently detected by the C1 complex (golden hexameric structure), which gets activated and cleaves both C2 and C4 [Merle *et al.* 2015a]. C4b and C2a assemble to the CP C3 convertase, which like the alternative pathway (AP) C3 convertase irreversibly cleaves a number of C3 molecules. This finally starts up the C3 amplification loop (red area), like described in Fig. 25. Both CP and AP C3 convertases bind free C3b, resulting in both C4bC2aC3b (the C5 convertase of the CP) and C3bBbC3b (the C5 convertase of the AP), which finally trigger the immunologically effective terminal pathway. Again, iC3b, but also iC4a, offer a permanent exit from the reaction chains, effectively supporting down-regulation.
When considering Cetuximab as a therapeutic antibody of the IgG type, it must be suggested that the C1 complex binds to it, triggering the complement CP wherever it binds to EGFR (comp. green antibodies in figure). While at the same time the sensitivity of the subsequent C3 amplification loop is increased due to Cetuximabs increasing effect on C3 itself, the C5 convertase-mediated activation of the terminal pathway should be considered. Consequently, acute inflammation should be suggested to occur at EGFR-expressing tissues (skin, colon, tumor). Again, given a suboptimally working C3 molecule due to the rs2287845 variant in several SR-negative patients, the effects of Cetuximab would be attenuated, as the CP triggers get not as strongly amplified like with wildtype C3 (no rash, no effectiveness against the tumor). Illustration and general mechanistic description following [Zipfel 2001] and G. Hegasy[22].

[22] http://www.hegasy.de (Downloads > Komplement System)

Taken together, the detected C3 variant offers a chance to generate a mechanistic model, potentially explaining the difference between patients being positive or negative for predictive skin rash upon Cetuximab treatment. The model (Fig. 27) includes two effects of Cetuximab: an increased C3 concentration (promoting both the alternative pathway initiation and the amplification loop) as well as the antibody's activation of the classical pathway. Consequently, the probability for the subsequent terminal pathway of the complement system to switch over to inflammation increases. As local tissue changes occur for the majority of patients, this leads to an overreaction in the skin (rash) and increased surveillance/destruction of the colonic tumor (survival).

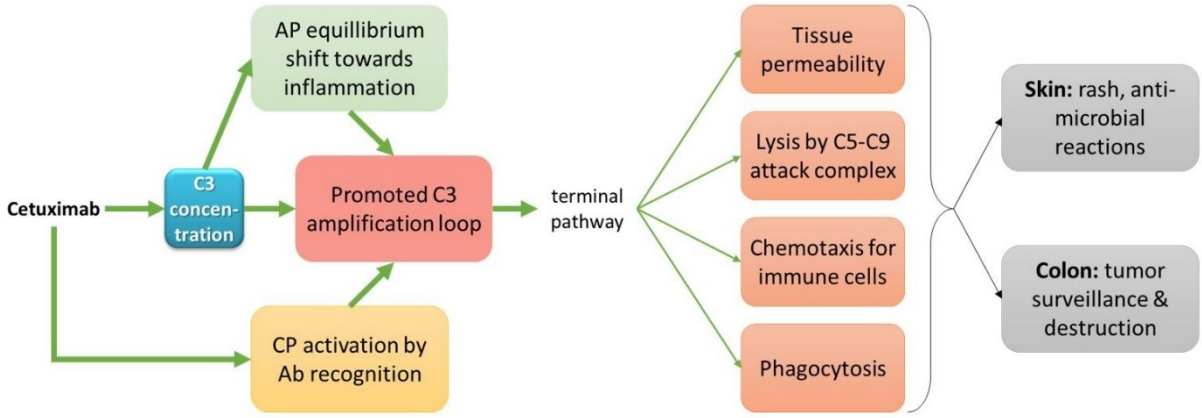


Fig. 27 - Overall model of C3-centric Cetuximab mechanisms causing rash in SR-positive patients. Cetuximab application triggers both the activation of the Classical Pathway (CP; yellow box) of the complement system and higher expression of the complement factor C3. The latter potentially promotes both the (constantly active) alternative pathway (AP; green box) and the amplification loop (red box). In sum, the gradual signal feeding the terminal pathway is elevated compared to no application of Cetuximab, leading to the activation of the terminal pathway. This switches the whole system to inflammation (orange boxes), causing the tissue effects in skin and colon/tumor (grey boxes).

In contrast, for SR-negative patients with rs2287845, C3 is altered and although principally functional, considered to be less active at least in the amplification loop. Speaking of the switch from a balanced steady state to massive immunologic cascades forcing local anti-tumor effects, the probability might consequently be lowered overall, leaving those patients less sensitive to associated immune triggers as Cetuximab in the given setting. Higher levels of (slightly impaired) C3 may be necessary to trigger both the adverse effect (skin rash) as well as the anti-tumoral effects, leaving Cetuximab with an undesirably low efficacy at usual therapeutical dosage levels. A similarely delayed signal switch upon an increasing gradual signal has already been simulated for the MAPK signaling cascade [Veitia 2004]).

Interestingly, the model of Cetuximab's effects via the complement system simultaneously covers the observation by [Chung *et al.* 2005] that Cetuximab therapy may also show anti-cancer activity in those patients whose tumors do not express EGFR. Also in those cases, although for the lack of molecular

targets the antibody recognition by the classical pathway may fail, the antibody's activating effect on the complement system at least via the alternative pathway would remain – given C3 is fully functional.

4.3.2 CCNK

CCNK codes for a member of the transcription cyclin family, cyclin K (CycK). Cyclins in general regulate the cell cycle both positively and negatively by interacting with Cyclin-dependent Kinases (CDKs). For a subgroup of CDKs, transcription-regulating functionalities have been described [Bartkowiak & Greenleaf 2011]. These are facilitated by phosphorylation of the C-terminal domain (CTD) of RNA polymerase II (Pol II). For CycK, two binding partners are known, namely CDK12 and CDK13 [Bartkowiak & Greenleaf 2011; Kohoutek & Blazek 2012]. While for CDK13 there is little specific knowledge, the Cyclin K/Cdk12 complex has been described to maintain genomic stability via the regulation of expression of DNA damage response genes [Blazek *et al.* 2011]. Also, central functions during development are controlled by the complex. Moreover, links to tumorigenesis have been drawn: in high-grade serous ovarian cancer (HGSOC), loss-of-function mutations of CDK12 have been described, predominantly leading to an inability to properly bind CycK [Ekumi *et al.* 2015; Joshi *et al.* 2014], ultimately leading to genomic instability. Meanwhile, in HER2-amplified breast tumors, overexpression of CDK12 and an association with high tumor grades occurs, offering a dual role as both tumor suppressor and oncogene for CDK12, and making it a potential pharmacological target [Paculová & Kohoutek 2017]. Cyclin K itself has been described as oncogenic in myeloma, with inhibitory effects on cell migration [Marsaud *et al.* 2010], and both proliferative and anti-apoptopic in prostate cancer, associating with poor recurrence-free survival [Schecher *et al.* 2017]. In terms of regulation, CycK has been suggested as a direct transcriptional target of the p53 tumor suppressor [Mori *et al.* 2002]. Regarding anti-cancer treatments, cyclin K expression has been detected to be down-regulated upon application of NS-398, a nonsteroidal anti-inflammatory drug associated with decreasing mortality in colorectal cancer [Zhang & DuBois 2001].

In terms of variation in the given dataset, CCNK shows with rs2069492 a quite common SNP (MAF_{1KG} = 0.447), although homozygously and exclusively assigned to the SR-positive group. Being located immediately upstream of exon 3, a splice region is annotated to be affected. Taking the interaction with CDK12 to be crucial for cell viability, slight variations in binding capabilities of cycK might predispose to human disease, as suggested by Bösken *et al.* 2014.

From the MSM's perspective, CCNK is connected to EGFR exclusively via SMAD2/3/4, each edge annotated as 'catalyze'. SMADs are direct downstream components of TGF-β signaling, which has been described to interplay with RTKs like EGFR in tumor development [Shi & Chen 2017]. However,

PubMed[23] does not reveal provide any evidence for CCNK's connection to SMADs. Moreover, neither CDK12 nor CDK13 are part of the MSM. While interactions to these two proteins have recently been added to Reactome according to the web portal[24,25], they were not listed in the functional interaction data provided by the ReactomeFI plugin within Cytoscape at the time of building the MSM. Functional interactions are updated once a year [Wu *et al.* 2014].

Taken together, the rs2069492 variant in CCNK might be a candidate of interest to check for in more patients, but should be regarded rather as a factor than a cause for the SR phenomenon.

4.3.3 CD86

CD86 (or B7-2), expressed by antigen-presenting cells, is a co-stimulatory transmembrane molecule regulating the balance between anergy and immunity of T lymphocytes. Known binding partners are CD28 and CTLA-4 [Pentcheva-Hoang *et al.* 2004], the latter being primarely expressed in CD8+ (cytotoxic) T cells. In cancer settings, CTLA-4 has an important role when considering immune evasion in the tumor's microenvironment [Webb *et al.* 2017], as it mediates the activation of T effector cells by negatively regulating their proliferation [Brunner *et al.* 1999]. While in EGFR-mutant non-small cell lung cancer CTLA-4 levels were negatively associated with overall survival [Soo *et al.* 2017], for metastatic melanoma an FDA-approved anti-CTLA-4 antibody, Ipilimumab, is available [Lipson & Drake 2011]. For this drug, similar to EGFRIs, skin rash was reported as a possible adverse effect [Dika *et al.* 2017]. Regarding genomic variants, with rs17281995 a UTR-3' region SNP has been described altering the binding of five different miRNAs, effectively influencing post-transcriptional regulation [Landi *et al.* 2012].

In the given Cetuximab-treated patient cohort, with rs2681417 a missense variant has been detected to be homozygously unbalanced between the patient groups. The variant's alleles refer to isoleucin or valin in the protein sequence, both being unpolar amino acids at slightly different sizes and structure in CD86's immunoglobulin domain. Accordingly, the protein's 3D structure might be varied, and as the domain facilitates binding functions [Bork *et al.* 1994], a change in strength of binding to CTLA-4 or CD28 appears possible. Ultimately, this would affect the regulative balance of T cell anergy and activation. Similar to other candidate genes, e.g. C3 or P3H3 (comp. Subsections 4.3.1 and 4.3.8, respectively), these regulative differences according to genetic variation might not be striking under physiological conditions, while under cancer treatment, balance outcomes might be discriminative.

[23] https://www.ncbi.nlm.nih.gov/pubmed
[24] http://www.reactome.org/content/detail/R-HSA-6797090
[25] http://www.reactome.org/content/detail/R-HSA-6797100

In terms of SR group assignments, the alternate allele appears to be the most common one. In turn, the minor, reference allele with an allele frequency of 0.132, is assigned to the SR-negative group. Assuming the minor allele is associated with a balance shift to anergy, both observations as well as the reduced anti-tumor efficacy in this patient group could be supported. Cytotoxic T cells are a central resource of inflammation (reduced or missing skin rash), and in the cancer context, they are able to identify and kill tumor cells [Martínez-Lostao *et al.* 2015], designating them to be an important biomarker in cancer research, especially in CRC [Deschoolmeester *et al.* 2010].

Taken together, CD86 deserves further study, presumably with immunostaining experiments or approaches on the composition and activity of cytotoxic T cells.

4.3.4 CDH11

Cadherin-11 (CDH11) is a cell adhesion molecule regulated by the Wnt pathway [Li *et al.* 2012] and TGF-β [Torres *et al.* 2013]. In turn, it regulates collagen and elastin synthesis [Row *et al.* 2016]. Although mainly expressed in bones[26] and brain[27], but also aorta, bladder and skin [Row *et al.* 2016], several observations for CDH11 have been observed in cancer settings. Several reports describe epigenetic silencing of CDH11 in metastasizing cancers (e.g. [Carmona *et al.* 2012]). While CDH11 has been described as pro-apoptotic tumor suppressor [Li *et al.* 2012] and for osteosarcoma patients, CDH11 expression was associated with survival [Nakajima *et al.* 2008], contradictory reports also exist. In CRC, CDH11 expression was markly increased in tumor tissue [Bujko *et al.* 2015], and in terms of statistic evaluation, a correlation with tumor invasiveness has been published very recently [Zhu *et al.* 2018]. Overall, the reports indicate for a dependency on the tissue type or other molecular factors like CALU [Torres *et al.* 2013], but leave the big picture contradictory.

Mechanistically, no direct association with EGFR signaling could be found: neither the literature, nor the MSM revealed a functional connection. When reviewing the shortest path between CDH11 and EGFR, ubiquitin-driven protein degradation turns out to be the connective element.

The variant c.1319C>T itself has been annotated as being a p.255P>M missense event. Regarding potential functional impacts on the protein level, it is located in an extracelluar cadherin domain, but without immediately influencing the Ca^{2+}-binding site amino acid positions[28].

[26] http://www.genecards.org/cgi-bin/carddisp.pl?gene=CDH11
[27] http://www.uniprot.org/uniprot/P55287
[28] https://www.ncbi.nlm.nih.gov/protein/NP_001788.2

Summing up, the findings leave no space for a mechanistic model linking the detected SR-negative associated CDH11 variant to the skin rash phenomenon.

4.3.5 COL4A4

COL4A4 codes for one of the six known chains of the collagen IV multi-meric molecule, which is the major structural component of the basement membranes ([Kalluri 2003]; for general details on collagens refer to Subsection 4.3.8 and Fig. 31) and is associated with diseases of the eye and the kidneys[29]. In terms of cancer, collagen IV plays a central role in the tumor microenvironment, providing signals regulating tumor growth and metastasis [Tanjore & Kalluri 2006]. In the same journal issue, Ikeda *et al.* reported expression levels of COL4A5/COL4A6 to be decreased in tumors by hypermethylation of the promotor regions, effecting a disrupted basement membrane.

For the coding variation position encoded by rs1800517, four alleles are known. Again, the reference allele is not the major one. In fact, the alternate allele 'A', detected in all SR-positive patients except one, is most common with a MAF of 0.547. For rs1800517, a clinVar entry exists, designating the alternate allele to be the benign one. The associated disease is the Alport syndrome, affecting the eye.

Anyhow, for none of the patients, relevant alterations of eyes or kidneys were reported. Also, several of the patients were detected with heterozygous variants, relativizing effects if given. Homozygous patients could be located in both groups. Together with lacking evidence for connections of rs1800517 to inflammation or cancer, COL4A4 was considered to be no promising candidate for further investigations in the given setting.

4.3.6 GRIP2

GRIP2, assigned to neurons as part of a multiprotein signaling complex[30], could not be related to either inflammation or cancer in a functional or mechanistic way. Anyhow, one report on prediction of colon cancer recurrence and prognosis stated GRIP2 as part of a signature of 15 differentially expressed genes, derived from a support vector machine (SVM) approach [Xu *et al.* 2017].

Regarding the detected variant rs9845816, dbSNP's web interface declares its effect on the protein level to be silent, as both reference and alternate allele code for the same amino acid. This is in contrast

[29] https://www.genecards.org/cgi-bin/carddisp.pl?gene=COL4A4
[30] https://www.genecards.org/cgi-bin/carddisp.pl?gene=GRIP2

to the results calculated offline (stating missense effects) and assumed to be according to recent reference sequence updates.

Taken together, GRIP2 appears of limited interest for further investigations.

4.3.7 NUP210

NUP210 encodes the nuclear pore glycoprotein 210 (Nucleoporin 210 kDa or gp210), which is a major structural compound of the nuclear pore complex [Greber *et al.* 1990]. NUP210 gets phosphorylated in the course of mitosis, when the pore complex gets disassembled. In the MSM, this is depicted e.g. by the edge to CDC42, which is on a shortest path to EGFR.

Although no mechanism is known yet, NUP210 has been reported to occur in risk studies on CRC [Landi *et al.* 2012; Cipollini *et al.* 2014], expression profiling of cervical cancer [Rajkumar *et al.* 2011] and an RNAseq analysis of lung adenocarcinoma [Kikutake & Yahara 2016], indicating a role in tumorigenesis. In the CRC-related studies, with rs354476 a UTR-3' variant has been described, in which the reference allele 'T' is recorded to be reduced expression of the gene if homozygous. In turn, the alternate 'A' allele is associated with an increased CRC risk. The authors draw the connection to post-transcriptional regulation via miRNAs, which hybridize with subsequences of the UTR 3' region, counteracting the translation to proteins. Interestingly, in primary biliary cirrhosis (PBC), a significant association could be drawn between anti-gp210 antibody production and a SNP in CTL-4 [Aiba *et al.* 2011], which is a binding partner of CD86 (comp. Subsection 4.3.3) and therefore potentially links NUP210 to T cell immunity.

In the given SR use case, four imbalanced variants could be found in NUP210 (Fig. 28). Among them, rs354478 is the only heterozygous one. Here, the reference, but minor allele 'C', is assigned to the SR-negative patient group. Although annotated to trigger a missense mutation p.1787Val>Met, both amino acids are unpolar, presumably not disturbing the local transmembrane region[31] of the protein. All three homozygous variants (rs2280084, rs2271504, rs2271509) occur exclusively in the same SR-negative patients, although showing distances of around 10 kB to each other. This makes them independent from individual short reads. Anyhow, just rs2280084 is declared to be a missense variant (Arg786Leu).

[31] http://www.uniprot.org/uniprot/Q8TEM1#subcellular_location

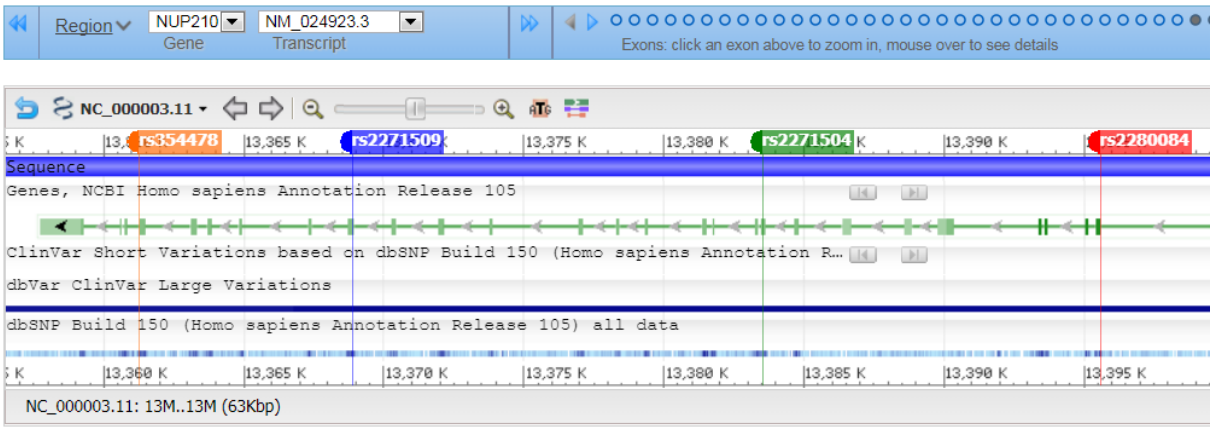


Fig. 28 - Overview on imbalanced variants in NUP210. Except rs354478 (orange), all variants occur homozygously and in the same patients. As NUP210 is located on the reverse strand of chromosome 3, the first variant in order is rs2280084 (red) in the beginning of exon 17. While rs2271504 (green) and rs2271509 (blue) are annotated to cause synonymous amino acid exchanges, the other two variants are labeled to be missense. For details compare Tab. 12 and Tab. 14. Screenshot from NCBI's Genome Data Viewer[32].

Anyhow, considering MAFs, impacts or patient group assignments (rs2280084 and rs2271504 both appear with the minor allele as being reference), in terms of causing the SR phenomenon all these candidate SNPs appear of limited interest.

Consequently, NUP210, although tagged with several variants (also comp. Tab. 10), appears of no deeper interest.

4.3.8 P3H3

Prolyl 3-hydroxylase 3 (P3H3; also known as LEPREL2 or GRCB) is an enzyme of the leprecan family, modifying collagen chains post-transcriptionally. It is part of a hydroxylation complex, which is located in the endoplasmatic reticulum (ER; [Heard *et al.* 2016]).

In breast cancer, P3H3 is a frequent target to epigenetic silencing by methylation and has been linked to carcinogenesis via association with higher tumor grades and Nottingham Prognostic Index [Shah *et al.* 2009]. On the one hand, this is in line with reports on silenced collagen-related genes, e.g. P3H2 [Shah *et al.* 2009], COL4A5/A6 (being major substrates the P3H family; [Ikeda *et al.* 2006]) and COL1A2 (the most frequent collagen; in primary CRC [Sengupta *et al.* 2003]).

In the given investigation, seven SR-negative patients were tagged with a frameshift mutation rs57050687 at the end of exon 1 (comp. Fig. 29). Consequently, major parts of the protein should be assumed to be dysfunctional.

[32] https://www.ncbi.nlm.nih.gov/genome/gdv/browser/?context=gene&acc=23225

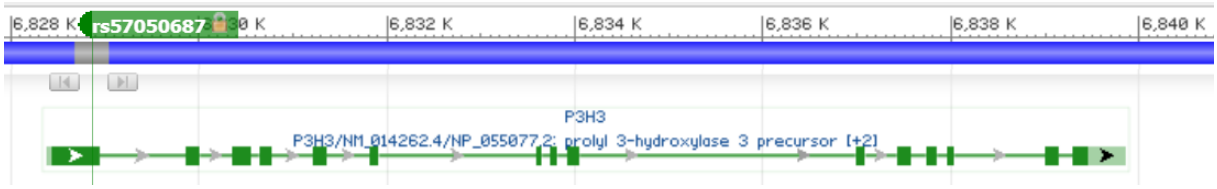


Fig. 29 - Detail view on rs57050687 in NCBI dbSNP (Genome Viewer). The green horizontal line depicts the mRNA of P3H3, with boxes representing exons. The green vertical line marks the position of the frameshift mutation rs57050687 at the end of exon 1, affecting the sequence to the right, probably hardly affecting the protein's amino acid sequence, structure and functionality. Viewer screenshot from dbSNP[33].

Confusingly, according to the Thousand Genomes Project (1KG as used in the pipeline) as well as ExAC[34], the references allele could not been observed in these cohorts[34], thus the MAF is 0 (and 1 for the alternate allele with an inserted 'G'). Only in GO-ESP[35], 22 individuals were detected with the particulary rare allele '-' (= no 'G'), resulting in a MAF of 0.0045. Although this may be due to the fact that 70 % to 80 % of the human reference genome originate from one individual [Osoegawa *et al.* 2001], in the given setting this implies that all SR-positive patients carry a quite rare variant. However, as coverage of the locus is low in the analysis, P3H3 is poorly studied and simultaneously highly interesting in the context of the use case, the gene was further studied, assuming a frameshift event, marking a difference between the SR patient groups.

Hydroxylation on pro-collagens is the most important modification during maturation, occurs on both proline and lysine residues after and even during translation, providing binding sites and consequently stability [Hudson *et al.* 2015]. This holds both for the formation of the quaternary structure of the collagen typical triple helix in fibrillar types (prolines) as well as for the higher order network-like assemblies in the extra-celluar matrix (ECM; lysines) [Hudson *et al.* 2015]. Overall, 28 different collagen types with various functions are known to date [Cescon *et al.* 2015]. Referring to this variety, the abundance (~30% of the whole-body protein content [Lehninger 1975]) and the complexity of proper maturation, a number of diseases is rooted in collagen-related pathology [Parvizi & Kim 2010].

Prolyl hydroxylases subdivide into two classes, depending on the targeted proline sites. While patterns of the form Gly-X-Y occur ubiquitously in collagens (with X und Y being any amino acid, but mostly proline), hydroxylases work either on the Y position (P4H) or the X position (P3H) and convert proline to 4-hydroxyproline (4Hyp) or 3-hydroxyproline (3Hyp), respectively. In quantitative terms, one in four amino acids in the ~1000 residue long collagen chains is a proline residue, of which about 40% are 4-

[33] https://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=57050687
[34] http://exac.broadinstitute.org
[35] http://evs.gs.washington.edu/EVS/

hydroxylated, but just up to 10 sites are 3-hydroxylated (depending on collagen type) [Hudson *et al.* 2015]. However, at least the latter is found to be highly conserved in the whole animal kingdom [Hudson & Eyre 2013; Weis *et al.* 2010].

Unfortunately, in comparison to P4H, the general role and particular function remains largely unknown for the P3H family [Weis *et al.* 2010]. The family overall consists of three homologous hydroxylases P3H1 (LEPRE), P3H2 (LEPREL1) and P3H3 (LEPREL2) plus two non-enzymatic, highly homologous paralogs CRTAP (LEPREL3) and SC65 (LEPREL4) [Hatzimichael *et al.* 2012; Gruenwald *et al.* 2014]. P3H1 (LEPRE1) has been linked to the specific 3-hydroxylation of Pro986 in collagen type I chains and, in the case of biallelic mutations in either P3H1 itself or its complex partners CRTAP and CypB, with recessive forms of osteogenesis imperfecta [Marini *et al.* 2014]. In contrast, P3H2 targets multiple sites in at least collagen types I, II, IV and V [Hudson *et al.* 2015]. Here, in terms of disease, connections to myopia have been drawn (e.g. in [Mordechai *et al.* 2011]). Solely for P3H3, substrate as well as disease-associated mutations still remain unknown [Hudson *et al.* 2017]. Meanwhile, at least the P3H3 complex formation with SC65 and, potentially, CypB and lysyl hydroxylase-1 (LH1) has been discovered [Hudson *et al.* 2017]. While CypB is the same protein like in the P3H1 complex, the reported interaction with LH1 revealed an interesting relation between the two most common post-translational modifications of collagens: primarely for $SC65^{-/-}$, but also for $P3H3^{-/-}$ mice, the major physiological collagens (types I, II, IV and V; type III is known to contain no 3Hyp at any site [Weis *et al.* 2010]) were clearly under-hydroxylated on lysine residues - but not on any of the known proline 3-hydroxylation sites [Hudson *et al.* 2017]. The former observation effectively phenocopied the Ehlers-Danlos syndrome (EDS) type VIA (six A; caused by loss the PLOD1 gene encoding LH1 [Steinmann *et al.* 2003]). The helical domain cross-linking sites of collagen type I, especially in the skin, appeared to be under-hydroxylated. However, the skin's resulting lack in structural integrity could only be detected on lab methods, rather than on physical appearance compared to wild type mice. Interestingly, heterozygous $P3H3^{+/-}$ mice did not reflect the described phenotype [Hudson *et al.* 2017], indicating loss of function just for homozygous variations. Similarely, P3H3 variations detected in the given SR setting also appeared exclusively homozygous.

The much more interesting observation from the recent study of [Hudson *et al.* 2017]is that for the major collagens of types I, II, IV and V no change in proline 3-hydroxylation could be recorded, although for both $SC65^{-/-}$ and $P3H3^{-/-}$ mice the P3H3 complex could not have formed. In other words: on major collagens, the prolyl 3-hydroxylation 3 complex is essential for hydroxylation of lysyls, but not for prolyls. The authors provide two explanations: on the one hand, and in line with the suggestion on collagen-modifying proteins to provide a dual function as both enzymes and chaperones [Ishikawa *et al.* 2009], P3H3 may have lost its hydroxylation functionality in the course of evolution, retaining the

LH1-supporting chaperone functionality. On the other hand, simply the proline substrate of P3H3 is still unknown, as the number of collagens goes far beyond the tested major types.

At this point, the look back to the MSM turns out to be helpful. As depicted in Fig. 21 and Fig. 22, all shortest pathes of functional interactions from P3H3 to EGFR (the target of Cetuximab) run across nodes belonging to five different collagen types: IV (COL4A[1-5]), VI (COL6A[1-3,5,6]), IX (COL9A[1-3]), XVII (COL17A1) and XVIII (COL18A1). As type IV has already been excluded to be a target of P3H3, investigations could be concentrated to just four out of 23 remaining types of collagens.

Collagen type VI is unusal compared to other fibrillar types due to its intracellular complexation going beyond trimerization of the helix structure. Rather more, tetramers of heterotrimeric helices of chains α1 (COL6A1), α2 (COL6A2) and α3 (COL6A3) are generated in a highly regulated assembly sequence and secreted afterwards. In terms of function, collagen type VI has been described as mechanically relevant due to cross-linkage with various ECM proteins (perlecan, collagens type I, II, IV and other), but also important for tissue homeostasis, e.g. by interaction with several membrane receptors linked to intracellular signaling pathways [Cescon *et al.* 2015]. In a not yet fully understood tissue-specificity, α3(VI)'s position may be occupied by alternative chains α4 (COL6A4), α5 (COL6A5) and α6 (COL6A6), suggesting different roles of COLVI variants in distinct tissues [Gara *et al.* 2011; Cescon *et al.* 2015]. In the skin, where collagen VI is expressed by fibroblasts, it is present in the dermis, abundant in the basement membrane, but absent in the epidermis [Gara *et al.* 2011]. Although several skin related abnormalities have been described [Lettmann *et al.* 2014], none of them resembles the SR phenotype. In terms of cancer, a set of cytoprotective functions is interesting: inhibition of apoptosis and oxidative damage, regulation of autophagy and cell differentiation, maintanance of stemness, and finally promotion of tumor growth, cancer progression and drug resistance [Cescon *et al.* 2015; Chen *et al.* 2013a]. Taken together, collagen VI contributes to at least five of the cancer hallmarks [Hanahan & Weinberg 2011]. When considering collagen type VI as substrate for P3H3, at least for a mechanistic model explaining the SR phenomenon, the potential of the candidate from both the skin and tumor perspective is given. Anyhow, the available information is insufficient, as no connection could be drawn to either EGFR or Cetuximab: neither integrins (ITGB3, ITGAV), nor platelet-derived growth factor (PDGFA, PDGFB) offered further meaningful interpretations of connections to EGFR.

Collagen type IX has been described as distinct component of cartilage and expressed in joionts, spinal disks and the eye's vitreous body [Olsen 1997]. Therefore, it is also unsupportive for a model explaining the SR-phenomenon.

Collagen type XVIII is ubiquitously expressed in different basement membranes of the whole body [Seppinen & Pihlajaniemi 2011] and maintains its integrity [Bager & Karsdal 2016]. Collagen XVIII is the

prime example for a functional dualism: the intact form holding structural properties, while signaling capabilities come into play upon degradation [Bager & Karsdal 2016]. This second function is provided by the C-terminal fragment, endostatin. It can be proteolytically cleaved, revealing anti-angiogenic effects [Marneros & Olsen 2005] and an inhibition of tumor growth by restiction of proliferation and migration as well as induction of apoptosis in epithelial cells [Seppinen & Pihlajaniemi 2011]. Anyhow, in the course of SR model generation, collagen type XVIII turns out not to be supportive. On the one hand, the defects in this collagen type are associated with abnormalities of the eye [Bager & Karsdal 2016], while none of those have been reported for SR-negative patients. On the other hand, like for collagen type VI, no connection to EGFR or Cetuximab could be drawn.

Collagen type XVII, also known as BP180, BPA-2 or BPAG2, revealed to be the most interesting molecule for a model connecting altered P3H3 in the sense of the SR phenomenon. This is in a nutshell due to the abundance at the skin's basement membrane, a functional dualism offering a direct link between cell adhesion and immunity, and a bi-directional regulative connection with EGFR supported by literature.

The molecule itself is a type II transmembrane protein (Fig. 30A), spanning the lamina lucida and projecting into the lamina densa of the epidermal basement membrane [Ujiie *et al.* 2010]. Structurally, it is a major component of the hemidesmosome (HD) type I (Fig. 30B), but not in type II [Moilanen *et al.* 2015]) complex, and crucial for stable adhesion of dermis and epidermis.

For collagen XVII, no investigations on 3Hyp sites have yet been reported. Anyhow, two closely similar and conserved motifs have been published from the collagen α1(I) chain (GLPGPIG**P**PGPR; hydroxylated proline highlighted) and collagen type II (G/PGPIG**P**PGPR; L→I conversion in italics) [Weis *et al.* 2010]. A BLAST search on the amino acid sequence of collagen XVII using these motifs revealed some imperfect hits (Tab. A1), providing hints on where such motifs could occur in the collagenous domains (Box A1). In any case, published knowledge on 3Hyp and its mechanisms is sparse, leaving space for experiments investigating on the yet hypothetical modification of collagen XVII by P3H3.

Defects in the HD proteins are associated with chronic skin fragility and blistering [Bruckner-Tuderman & Has 2012]. In turn, for e.g. tissue repair and immune cell invasion, but also keratinocyte migration, disassembly of hemidesmosomes is required [Löffek *et al.* 2014]. Growth factors like EGF are capable of triggering this effect by phosphorylating α6β4 integrin [Margadant *et al.* 2008]. As disassembly is mostly transient, HD stability has been suggested to be viewed as a dynamic equilibrium, with its components (primarely α6β4 integrin) in rapid turnover [Wells 2006]. In cancer settings, this equilibrium may be shifted towards migration, facilitating tissue invasion and metastasis [Löffek *et al.* 2014].

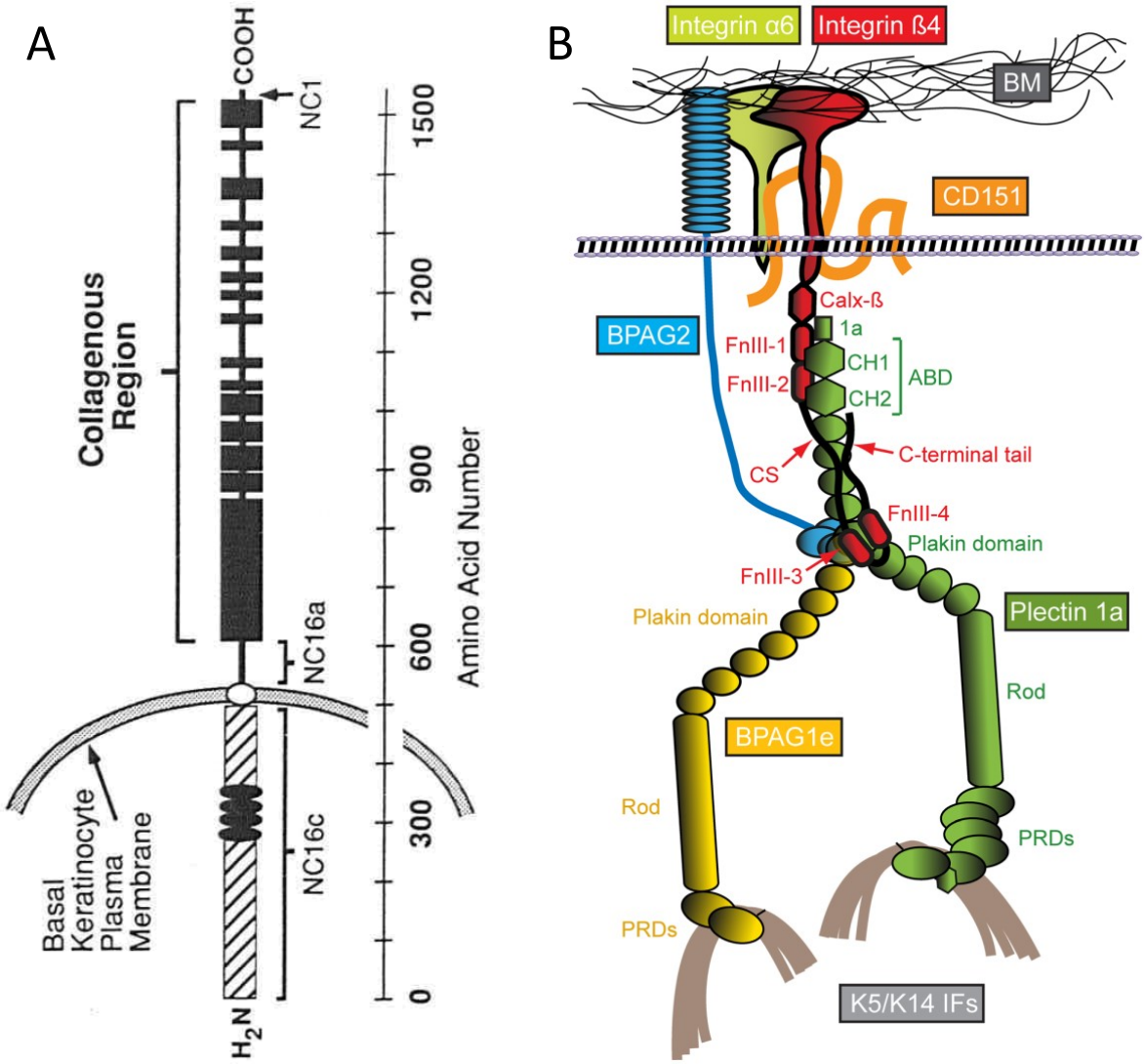


Fig. 30 - Collagen XVII molecular structure and as component of the hemidesmosome. A: Collagen XVII (according to [Giudice *et al.* 1992]; there named BP180) as type II transmembrane protein consists of a C-terminal extracellular part, one trans-membrane domain and a N-terminal intracellular part. The authors numbered 15 collegenous domains (COL1 to COL15) starting from the C terminus, interrupted by 14 numbered non-collagenous (NC) domains. The last domain NC16 divides into the sub-domains a (juxta-membraneous), b (membrane-spanning) and c (containing four tandem repeats). **B:** Classical type I hemidesmosomes (according to [Walko *et al.* 2015]) are cell-matrix junction complexes, besides collagen XVII/BPAG2 consisting of transmembrane α6β4 integrin and CD151, intracellular BP230/BPAG1e and plectin. While the extracelluar components bind to laminin-332 and collagen IV [Nishie *et al.* 2011] in the basement membrane (BM), the intracellular side binds to the intermediate filament (IF) cytoskeleton (Keratins K5 and K14).

As indicated by the MSM, the connecting nodes between EGFR and collagen type XVII are the closely related sheddases ADAM10 and ADAM17 (comp. Fig. 21, Fig. 22), which generally show overlapping substrate specificity [Gooz 2010]. Interestingly, this connection discriminates the COLXVII node from all other shortest path nodes indicating for collagens of types IV, VI, IX and XVIII: those connect via cell adhesion molecules (NCAM1, IGTB3, ITGAV) or the growth factors PDGF[A/B].

Sheddases, in particular ADAMs, both constantly and upon signaling, cleave various membrane-bound proteins on the extracellular domain, often initiating further reactions by the cleavage products [Giebeler & Zigrino 2016]. In fact, both collagen type XVII [Nishie *et al.* 2011] and membrane-bound EGF (e.g. amphiregulin by ADAM10 [Lao & Grady 2012], HB-EGF by ADAM17 [Yin & Yu 2009]) are constitutively shed from the cell surface in physiological processes and thereby released into the intracellular space. Interestingly, in particular levels of ADAM17 are elevated in human colorectal tumors [Arribas & Esselens 2009], designating it as target for combination therapies with EGFRIs [McGowan *et al.* 2013]. While ADAMs act on the extracellular domains of proteins, it can assumed that collagen type XVII molecules bound to hemisdesmosomal complexes are sterically inaccessible and consequently uncleavable. Similarely, lipid rafts seem to play a role in regulating shedding, as collagen XVII and α6β4 integrin as HD components locate to those special areas of the cell membrane [Zimina *et al.* 2005], while ADAM10 and ADAM17 are excluded [Murai 2012]. While sheddase activity on EGF family members trans-activates the EGFR and the subsequent pathway down to growth [Tanida *et al.* 2004], shedding of collagen type XVII triggers its secondary function. Like described for collagen XVIII above, collagen XVII comes with a dualism beyond mechanic stability. In fact, the shedding of extracellular domains (ecto-COLXVII; [Nishie *et al.* 2011]) triggers local immune reactions (complement activation, mast cell degranulation and neutrophil infiltration [Panelius & Meri 2015]) as well as cell migration [Parikka *et al.* 2003]. Besides the physiological fragment, at least two disease-related ones are known (LAD-1, LABD97; [Hirako *et al.* 2003]). Although antibodies are assumed, the binding partner of the soluble fragment in terms of immunologic reactions remains unknown [Nishie *et al.* 2011; Nishie *et al.* 2015; Moilanen *et al.* 2015].

Thus, the hemidesmosomal form might be referred to as adhesion-related, while the unbound form is more dedicated to signaling. From a tissue point of view this makes sense, as inflammation goes along with decreased cell-cell adhesion, providing a tissue structure permissive for recruited immune cells. The other way round, healthy tissue in homeostasis benefits from tight cell-cell interactions, especially for those layers intended to be a impermeable layer to ‘non-self’ (i.e., the skin’s microbiome).

Immunologically, collagen XVII has been reported to have an attenuating effect on interleukin 8 (IL-8) via NFκB [van den Bergh *et al.* 2012]. IL-8 is an immunological key factor, promoting neutrophil recruitment, angiogenesis and inflammation [Janeway 2001], but also growth, e.g. by trans-activating

EGFR trough ADAM10 [Tanida *et al.* 2004]. This trans-activation mechanism in fact is induced by IL-8 itself [Itoh *et al.* 2005], offering a signaling link between collagen XVII and EGFR as indicated by the functional interaction data behind the MSM. To this point, the MSM could only deliver statements on independent, pairwise interactions between molecules of the given shortest pathes, not necessarely acting together in one scenario.

Taken together, for P3H3 a hypothetical model of action could be constructed, potentially explaining the SR phenomenon in Cetuximab-treated CRC patients (Fig. 31).

In the skin of untreated individuals with wildtype P3H3, collagen XVII gets regularely modified on its secretion to the cell membrane. EFGR pathway activity is well balanced and decreases with the keratinocytes developmental progress, during which they undergo maturation on the way from the skin's basal membrane to the body surface [Candi *et al.* 2005]. With progress in maturation, the grade of differentiation increases, and motility decreases [Fuchs & Raghavan 2002]. Regarding the balance between collagen XVII's oposing states, EGFR triggers hemidesmosomal disassembly preferrably in early stages, promoting cell migration [Walko *et al.* 2015]. The weaker EGFR signaling gets, the more the balance of collagen XVII states shift to adhesion via assembly of hemidesmosomes, triggering linkage to collagen IV and laminin of the intercellular space. Simultaneously, growth signaling is attenuated by a negative feedback inhibition on the downstream pathway being linked to collagen XVII's state (HD-bound or not), which may be related to the increasing inhibition of IL-8 via NFκB, effectively lowering the activity of sheddases [van den Bergh *et al.* 2012].

In a Cetuximab treatment of P3H3 wildtype individuals, the systemically applied therapeutic antibody reaches the skin and silences EGFR, consequently disturbing the dynamic equillibrium of skin maturation processes. Molecularely, in such a setting the balance of collagen XVII states would be affected, shifting the balance in favor of the hemidesmosomal conformation, leading to a premature differentiation of keratinocytes. Consequently, the skin barrier gets impaired, ingesting bacteria from the skin surface into deeper layers [Holcmann & Sibilia 2015]. Here, immune cells, but also epithelial cells, detect microbe-associated molecular patterns (MAMPs) like lipopolysaccharides (LPS; by TLR4 [Ray *et al.* 2013]) or flagellin (by TLR5 [Miller *et al.* 2005]). Recognition of those molecules triggers proinflammatory immune responses e.g. via interleukin (IL-)8 expression [van den Bergh *et al.* 2012; Fraser-Pitt *et al.* 2011]. As many bacteria at a time infiltrate the skin, immune reactions can be expected to be strong, leading to local, visible inflammation, i.e. skin rash. Simultaneously, early and modest immune reactions invoking collagen XVII's shed extracellular domain are impaired, as sheddable collagen XVII molecules outside hemidesmosomes are rare.

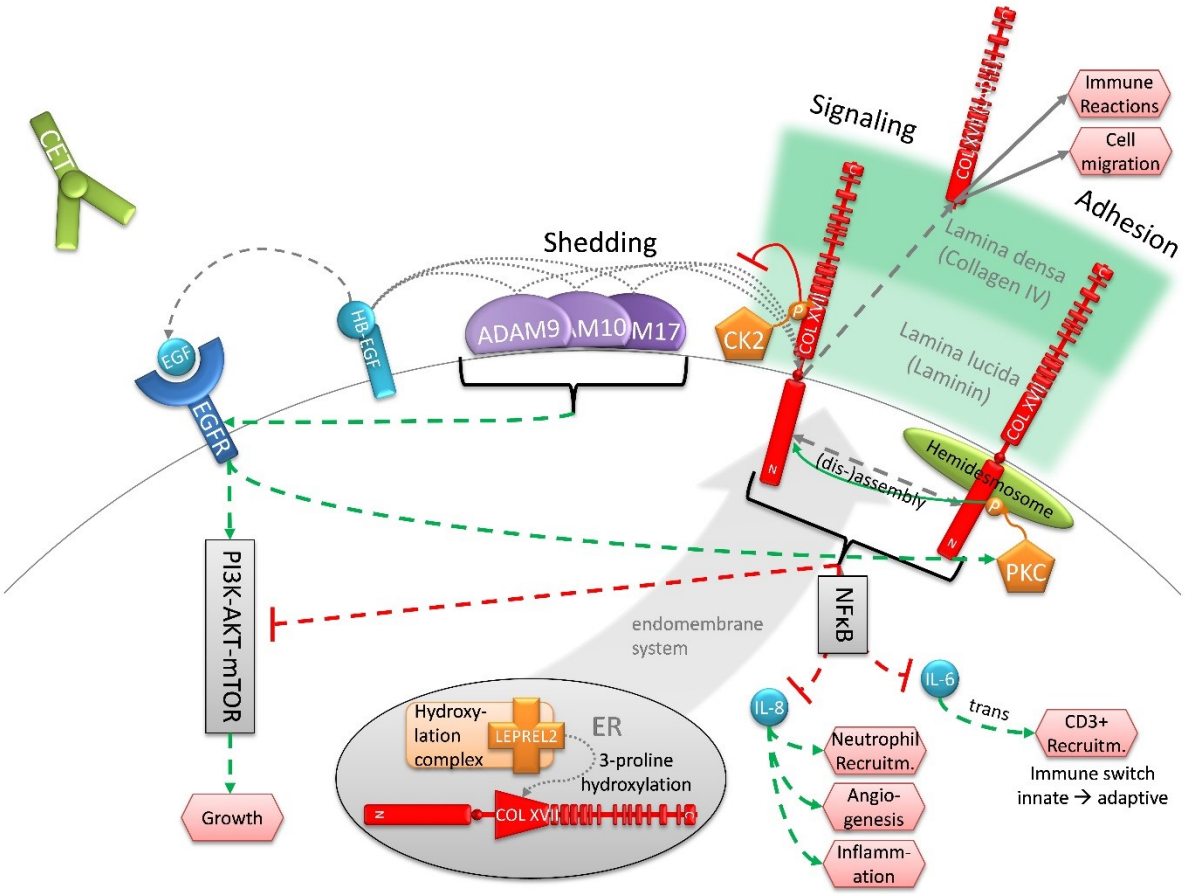


Fig. 31 - Model on P3H3, collagen XVII and ADAM-driven shedding mechanisms. Collagen XVII (red rod; for an amino acid-level depiction comp. Box A1), synthesized in the endoplasmatic reticulum, gets post-translationally modified by various enzymes, including P3H3, which is organized in a hydroxylation complex [Heard *et al.* 2016]. Through the endomembrane system it reaches the cell's outer membrane. The extracellular region, containing 15 collagenous domains, binds to collagen IV and laminin-332 [Nishie *et al.* 2011]. In a dynamic equilibrium the molecule toggles between it's adhesion-related, lipid raft-located, hemidesmosome-bound state (promoting keratinocyte differentiation) and an unbound state, which is signaling-related. Disassembly of the hemidesmosome (HD) is promoted by protein kinase C (PKC; [Kitajima *et al.* 1992]) and can be induced by growth factors like EGF [Oliva *et al.* 2005]. Upon leaving the lipid raft, collagen XVII gets exposed to ADAMs. If, ADAM10 or -17 (the later also named TACE) sheds the extracellular domain in the juxta-membranous region [Nishie *et al.* 2011], given that such cleavage is not hindered due to target phosphorylation by CK2 [Zimina *et al.* 2007]. Shedding enables pro-inflammatory signaling and promotes cell migration, as the molecule is not available for HD re-assembly anymore. Presumably, at least major parts of the observed immune reactions are based on release of IL-8, as collagen XVII's attenuating effect (through NFκB pathway [van den Bergh *et al.* 2012]) is lost. Also IL-8 [Itoh *et al.* 2005], and additionally EGFR ligands (HB-EGF, AREG; [Lao & Grady 2012; Yin & Yu 2009]), get shed by ADAMs, leading to activation of EGFR signaling, promoting growth and closing the cycle to PKC (leading to further HD disassembly). As the IL-8 receptor (IL8R), like several other G protein-coupled receptors (GPCRs; [Wang 2016]) transactivates EGFR via ERK1/2 and increased ADAM-related sheddase activity [Yin & Yu 2009], a second, potentially amplifying cycle may be considered here.

In the colon, collagen XVII is supposed to be generally more dedicated to signaling and cell migration rather than adhesion and tissue integrity [Moilanen *et al.* 2015], which would significantly support tissue invasion in the case of carcinoma cells. In fact, collagen XVII expression has been reported as significantly associated with higher TNM stage, decreased disease-free and cancer-specific survival [Moilanen *et al.* 2015]. Anyhow, EGFR signaling is expected to influence hemidesmosomal integrity, as activation also causes the cytoplasmic phosphorylation and dissociation of integrin α6β4 integrin, being part of the hemidesmosomal complex type II [Mariotti *et al.* 2001], which is found in simple epithelia of the intestine [Walko *et al.* 2015].

When considering SR-negative patients carrying the rs57050687 variant and consequently impaired P3H3, a suboptimal post-translational modification of COLXVII in might occur in terms of prolyl 3-hydroxylation. Unfortunately, collagen XVII has not yet been examined for such positions, designating it still as a hypothetical target to P3H3 (while to date not a single target is known [Gjaltema & Bank 2017]). In terms of proline residues suitable for 3-hydroxylation, literature revealed just two explicit similar P3H motifs being conserved across vertebrate species, although generalized evolutionary studies exist [Hudson *et al.* 2014]. In human collagen type I this is GLPGPIGPPGPR, and GIPGPIGPPGPR in human collagen type II [Weis *et al.* 2010]. For those, a BLAST search in collagen XVII's sequence ended in ten positions with an expect value <= 0.01 (comp. Tab. A1), distributed over ten collageneous domains (COL5 to COL15; comp. Box A1). Considering at least one position being fact, an altered hydroxylation pattern can be assumed in the case of defect P3H3, affecting the final conformation of the collagen XVII protein.

Taking into account the dual function of collagen XVII, adhesion and signaling, conformational changes might influence both. Referring to Fig. 31, two particular mechanisms may be affected: the assembly-disassembly switch at the hemidesmosome (HD), as well as the shedding of the extracellular membrane, in both cases due to steric hindering during the interaction with other proteins (integrin in HDs and ADAMs at shedding). Overall, the equilibrium between adhesion and pro-inflammatory signaling would be affected. As SR-negative patients show less inflammation in the skin, a shift to the adhesion state can be assumed – potentially due to impaired shedding, removing less molecules permanently from the 'pool' of those available for HD assembly. On the other hand, regarding HDs adhesion to molecules in the extracellular matrix (ECM), proper binding of collagen XVII to either collagen IV or laminin-332 [Nishie *et al.* 2011] appears questionable. Thinking of impaired expression of at least type IV collagen to be an early event for invasive phenotypes [Ikeda *et al.* 2006], altered tissue behaviour in the tumor and effects on survival should be considered. In terms of collagen XVII's immunological role, considered to be negatively influenced by P3H3 failure, tumor surveillance could

be simultaneously lowered. Consistently, direct anti-proliferative effects of active P3H3 have been reported at least in breast cancer settings [Shah *et al.* 2009].

From a clinical point of view, individuals with sub-optimally processed collagen XVII due to P3H3 failure would not necessarily show a phenotype in daily life. As reported for the P3H3$^{-/-}$ mouse model, skin outer appearance is unremarkable, although of reduced structural integrity upon mechanic lab examination [Hudson *et al.* 2017]. As stated earlier, mutations in P3H3 may be responsible for mild variants of the EDS VI [Heard *et al.* 2016]. For the SR-negative patients affected by the detected homozygous P3H3 mutation a simple clinical check for soft skin, lax joints and kyphoscoliosis may be helpful in order to proof or reject the hypothesis of P3H3 being causative. Also, mild changes in general hair development might be a visual marker ahead of therapy, as expression of collagen XVII is high in hair follicle stem cells [Gude 2011].

In a nutshell, P3H3 deserves deeper investigations, particularly in the light of overall sparse knowledge on this protein and its functions. As a key point occur HD assembly and disassembly, as their exact dynamics remain elusive [Walko *et al.* 2015].

4.3.9 STUB1

STUB1 (also known as CHIP or STIP1) is both a ubiquitin E3 ligase and co-chaperone, associated with heatshock proteins [Joshi *et al.* 2016]. Due to a variety of ubiquitylation targets there are several relations to biological mechanisms known, e.g. base-excision repair [Parsons *et al.* 2008], TGFβ signaling [Shang *et al.* 2014], T cell immunity activation [Chen *et al.* 2013b; Wang *et al.* 2013] and autophagy [Joshi *et al.* 2016]). Consequently, a role in multiple diseases has been described, e.g. ataxias [Heimdal *et al.* 2014], systemic lupus erythematosus [Guo *et al.* 2016]), Alzheimer's and Parkinson's disease [Joshi *et al.* 2016]. In terms of cancer, most prominent is STUB1's tumor-suppressing role, facilitated mainly through the LKB1-AMPK pathway [Gaude *et al.* 2012]. While LKB1 is a master kinase controlling metabolism, growth and cell polarity, it is known to be inactivated through mutation during carcinogenesis [Shackelford & Shaw 2009]. In colon cancer settings, STUB1 has also been reported to act as a tumor suppressor, repressing NFκB-mediated signaling [Wang *et al.* 2014b]. In contrast, STUB1 also enhances angiogenesis, at least after cardiac infarcts [Xu *et al.* 2013].

Most interestingly, a direct regulative link between STUB1 and EGFR has been reported for cell culture [Chung *et al.* 2016] as well as pancreatic cancer [Wang *et al.* 2014a]. In the cell culture experiments, overexpressed STUB1 turned out to specificially ubiquitinate mutant EGFR, resulting in proteasomal degradation of the targeted receptor as well as inhibition of cell proliferation and xenograft's tumor growth. As mutant EGFR is directly related to EGFR TKI resistance, the authors suggest STUB1 to be a

potential novel therapeutic target to overcome such settings. Without restricting these observations to mutant EGFR, in the pancreatic cancer setting the oberservations were made as well. Additionally, reduced tumor migration and invasion were reported. Not last, enhanced apoptosis induced by Erlotinib and a negative correlation of STUB1 expression with tumor differentiation and overall survival marked clinical significance. Similarely, in gastric cancer low STUB1 expression levels correlate with a clinically aggressive phenotype [Gan *et al.* 2012].

Although for the patients in the SR setting EGFR is given in its wildtype form, the detected variants indicate for structural variations of STUB1 (Fig. 32). Again, the reference but less common allele, is assigned to the SR-negative group. SNPeff annotates these variations to be located within a splice region. The gene's 3' end codes for a protein's carboxy terminus, which in the case of STUB1 is in fact the E3 ligase or U-box domain [Joshi *et al.* 2016], facilitating ubiquitylation and consequently protein degradation. An altered efficacy of STUB1 in consequence of slight structural modifications in this important domain should be considered.

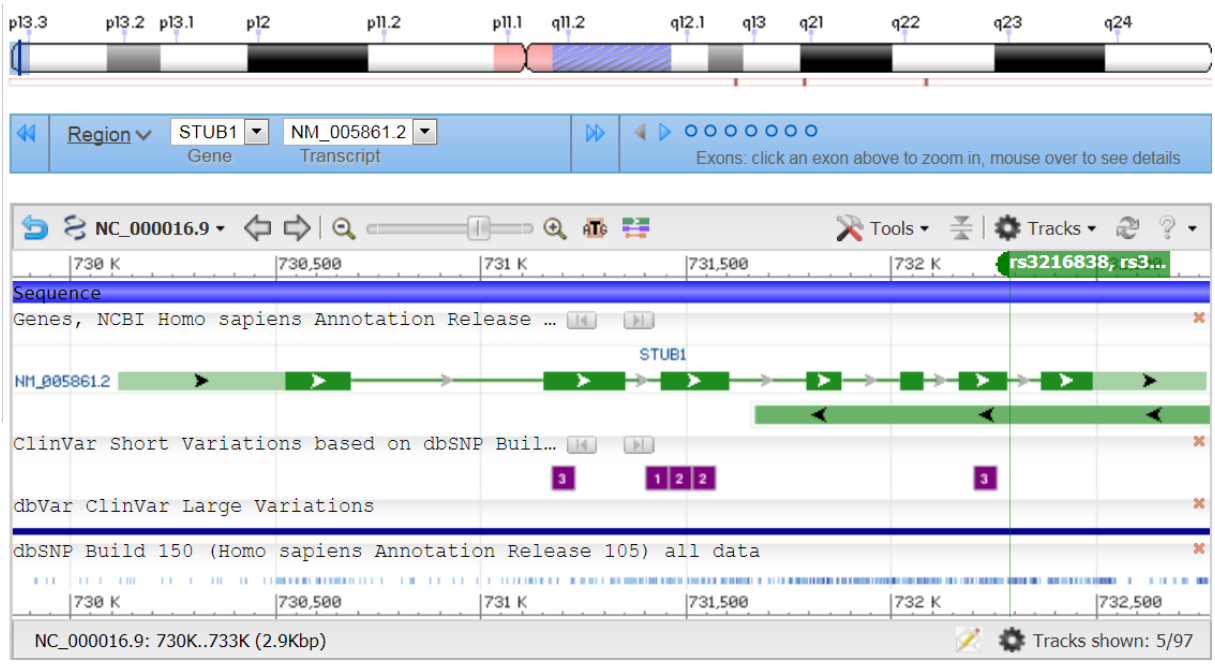


Fig. 32 - Location of detected SNPs in STUB1 gene structure. The position locates right after exon 6 in a stretch annotated to be a splice region. Screenshot from NCBI's Genome Data Viewer[36].

Taken together, investigating the action of STUB1 in the cell offers chances for closer insights into the SR phenomenon. Measurements of expression and activity levels in both SR-negative and -positive patients appear necessary.

[36] https://www.ncbi.nlm.nih.gov/genome/gdv/browser/?context=gene&acc=10273

4.3.10 TLR5

Toll-like receptor 5 (TLR5) is a member of group of pattern recognition receptors (PRRs), which being part of the innate immune response detect a variety of microbial- (or pathogen-)associated molecular patterns (MAMPs/PAMPs) [Köllisch *et al.* 2005]. In the case of TLR5 this is bacterial flagellin, which is quite ubiquitously present at environmental contact surfaces like skin and colon. Apart from immune reactions, TLR5 takes part in the induction of epithelial cell responses including cell migration, wound repair, proliferation and survival [Shaykhiev *et al.* 2008]. Transactivation of EGFR via the release of TGF-α by sheddases (comp. Subsection 4.3.8) hereby is a central mechanism [Yu *et al.* 2012]. In terms of cancer, TLR5 has been described to modulate tumor development (mouse xenograft model of human colon cancer; [Rhee *et al.* 2008]). In several studies an altered T cell-related immunity was observed upon TLR5 activation (e.g. [Ogino *et al.* 2013]), with some reporting a metastasis suppressive activity on metastasis [Brackett *et al.* 2016] as well as tumor rejection [Marshall *et al.* 2012]. Interestingly, in a 5-fluorouracil (5-FU) application setting, a reduction of systemic toxicity, selectively for non-tumor cells, has been reported for Entolimod (an TLR 5 agonist; colon adenocarcinoma mouse model; [Kojouharov *et al.* 2014]). Similarely, a bioinformatic study revealed TLR5 to be one of six genes whose expressions were closely related to responsiveness to oxaliplatin treatment [Klahan *et al.* 2016]. Both drugs were applied (5-FU as Capecitabine in the CIOX study), additionally to Cetuximab.

Several functional polymorphisms are known for TLR5, including connections to systemic lupus erythematosus (SLE; [Hawn *et al.* 2005]) and psoriasis [Loft *et al.* 2017]. The detected variation rs5744174 is not yet linked to a disease, but leads to a missense variant F616L, located witin the LRRCT region of the protein[37]. LRRCT is a binding interface domain intended to support dimerization of the receptor upon complexation with flagellin [Berglund *et al.* 2015]. Therefore, variations in this region potentially affect binding affinities and consequently downstream signaling. In the case of TLR5, besides EGFR this is such a downstream element, as well as the pro-inflammatory interleukin-8 chemokine (IL-8; [Fraser-Pitt *et al.* 2011]). In fact, those two interact widely independent, with IL-8 signaling being depended on the NFκB pathway [Gao *et al.* 2010]. Meanwhile, pro-inflammatory effects of TLR5 signaling are attenuated via a negative feedback loop; here, activated EGFR phosphorylates MUC1, whose cytoplasmic tail consequently associates with TLR5 [Kato *et al.* 2016], effectively competing with My88 as downstream signaling partner [Kato *et al.* 2012].

Anyhow, the minor missense variant appears to be enriched in the SR-positive patient group. Considering the low fraction of SR-negative patients, the TLR5 variant reported here might not be considered as causative.

[37] http://www.uniprot.org/uniprot/O60602#family_and_domains

4.3.11 KISS1

After all, one candidate missed by the MSM remains worth shortly mentioning: KISS1. SNPeff reported with rs71745629 a frameshift variant (MAF_{1kG} = 0.222) due to an deleted 'T' (forward strand; 'A' in reading direction), which in dbSNP appeared immediately destructive for the stop codon at the end of the terminal exon 3 of the canonical transcript (p.139Ter>Trp). Consequently, the transcript would be elongated. In fact, although SNPeff reports one transcript, Ensemble knows two isoforms: KISS1-201[38] and a 21 bp longer form KISS1-202[39]. NCBI RefSeq[40] reports the common length of 138 amino acids for the fully translated protein, and also mentions the polymorphic site encoded by rs71745629. Furthermore, it is thereby stated that in the case of the absence of the stop codon a downstream one is utilized instead, extending the protein for additional seven amino acid residues. Interestingly, several authors refer to either Lee *et al.* 1996 or West *et al.* 1998, stating a length of 145 amino acids for the protein. This difference in length of seven amino acids implies that the earliest publications describing KISS1 studied the less frequent isoform. As anyhow the gene product of KISS1 is a precursor being cleaved into active peptides[41] (so-called kisspeptines (KP); comp. Fig. 33), which act as natural ligands for the G protein-coupled receptor GPR54 [Kotani *et al.* 2001].

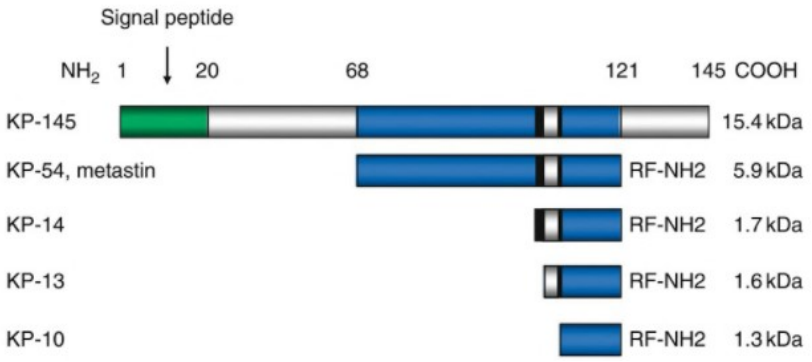


Fig. 33 - From KISS1 gene product to active peptides. The full length gene product has been described both to be composed of 138 and 145 amino acids, possibly referring two isoforms not being discussed in one common study yet. In any case, after cleavage of the full length precursor protein, KP-54 forms, with an amidated C-terminus at the RF motif (R = arginine, F = phenylalanine). All shorter peptides are formed by further cleavage, preserving the C-terminal end. A cleavage of the last three amino acids 119 to 121 by MMPs is possible, effectively inactivating the peptide [Takino *et al.* 2003]. Scheme from [Teles *et al.* 2011].

GPR54 (or KISS1R) appears overexpressed in triple negative breast cancer (no expression of estrogen receptor α, progesterone receptor and human epidermal growth factor receptor 2), promoting drug

[38] https://www.ensembl.org/Homo_sapiens/Transcript/Summary?db=core;g=ENSG00000170498;r=1:204190341-204196486;t=ENST00000367194
[39] https://www.ensembl.org/Homo_sapiens/Transcript/ProteinSummary?g=ENSG00000170498;r=1:204190462-204192876;t=ENST00000625357
[40] https://www.ncbi.nlm.nih.gov/protein/NP_002247.3
[41] http://www.uniprot.org/uniprot/Q15726#ptm_processing

resistance [Blake *et al.* 2017]. Several other authors, in contrast, reported the expression of the ligand precursor KISS1 to be an important anti-metastatic factor, setting disseminated cancer cells to a dormant state [Beck & Welch 2010], without affecting tumourigenicity or local invasiveness [Miele *et al.* 1996]. Complementary, methylation of KISS1 and decreased expression level correlate with higher tumor stages and grades, shorter survival and recurrence in colorectal cancer [Moya *et al.* 2013; Okugawa *et al.* 2013; Huo *et al.* 2018], but also other cancer entities [Tng 2015]. Anyhow, in further reports, e.g. for patients with colorectal adenocarcinoma, contradictory observations were reported: levels of KISS were higher in larger tumors and stage III/IV cancers, and patients with greater KISS1 levels had worse prognosis [Kostakis *et al.* 2015]. There, GPR54 could not be detected in either normal or the malignant colonic epithelial cells.

Functionally, for kisspeptine-expressing cells, both inhibition of chemotaxis and cell invasion could be observed [Mead *et al.* 2007], possibly mediated by downregulation of matrix metalloproteases (MMPs) by KPs [Hesling *et al.* 2004]. KP-driven regulation of MMPs may occur on both the transcriptional and protein level. While nuclear factor-κB binding to the MMP-9 promoter region has been reported to be reduced by KPs [Yan *et al.* 2001], stable complex formation has been described for the N-terminal 48 amino acids of the KISS1 full length protein product and pro-MMP-2 or pro-MMP-9, yet without further knowledge on the actual physiological consequences of this interaction [Mead *et al.* 2007]. In turn, active MMPs are able to cleave the bond between Gly118 and Leu119 in KPs, leading to inactivation of KP and potentially representing a regulatory feedback mechanism between KP and MMPs [Takino *et al.* 2003]. Downstream signaling of the KISS1 peptide receptor, GPR54, seems to be a separate way of action, involving numerous molecular factors (like ERK1/2, MAPKs, PKC, PLC and several more; comp. Mead *et al.* 2007), potentially explaining the various physiologial roles described for kisspeptines, which also include endocrinologic functions in the brain [Oakley *et al.* 2009], pregnancy [Mead *et al.* 2007] and puberty [Teles *et al.* 2011].

Interestingly, GPR54 has been reported to transactivate EGFR, promoting cell invasiveness in breast cancer [Zajac *et al.* 2011]. Transactivation of EGFR by peptide G protein-coupled receptors has been reported as important mechanism in cancer cell proliferation relatively recently [Wang 2016]. While agonists for such receptors may bypass EGFR blockade in EGFRI treatments, antagonists are supposed to potentially support treatment by increasing cytotoxicity of TKIs [Moody *et al.* 2016]. In fact, the clinical exploitation of KISS1 to treat metastases has already been claimed [Beck & Welch 2010].

However, neither for the skin nor for the colorectal tumors expression is yet known for GPR54. This anyhow leaves space for theories related to MMPs and modifications of the extracellular matrix. The core question is in fact, whether the additional seven amino acids of the less common KISS1 isoform shows a deviant mechanistic behaviour in cleavage or binding to either GPR54 or MMP precursors.

Although speculative, the mRNA of the less common might not be less stable, as the last exon also encodes the polyadenylation signal [Harms & Welch 2002].

In a nutshell, in terms of cancer mechanisms and epidemiology, KISS1 appears as interesting as contradictory, with its exact systemic modes of action remaining elusive.

4.3.12 Further candidates and summary

Due to the MSM filtering effect, some candidates from the initial list of SR-positive associated variants (comp. Tab. 9 and Tab. 10) have been cut. Among them several appear anyhow of low potential: for **HNRNPK** (Heterogeneous Nuclear Ribonucleoprotein K), an exceptional number of synonymous variations were detected in exons 9 and 13, of which only one could be linked to a dbSNP ID (rs code). Neither the gene, nor associated disease or predicted mechanistic variation effects offered a link to the SR phenomenon. Similarely, for **CAPN11** (calpain 11, a protease; 3' UTR variant) and **MCPH1** (microcephalin, a DNA damage response protein during cell cycle; missense event leading to a p.875Pro>Leu substitution) no links could be established.

BAIAP3 (BAI1 associated protein) encodes a gene targeted by p53, has been described as a brain-specific angiogenesis inhibitor and is associated with the KEGG pathway 'Transcriptional misregulation in cancer'. The detected splice site variant rs2235632 might prevent splicing out during transcript maturation, but a mechanistic role for the clinical SR setting remains elusive. **CRB2** (Crumbs 2, a cell polarity complex component), shows a missense event in exon 3, coding for a calcium-binding EGF-like domain[42]. However, an exchange of the two smallest amino acids p.159Gly>Ala may be considered to be largely without an effect. For this gene, no connection in terms of disease or general regulation could drawn.

Considering those variants being linked to gene represented in the MSM, two appeared just being low impact (intronic; in **CASK** and **MARK3**). Even a medium impact, like the missense events in **OBSCN** (obscurin), does not imply a connection to the clinical seeting: OBSCN is related to muscular tissues.

In a nutshell, these other candidates do not appear promising for further investigations. Accordingly, CCNK, CDH11, COL4A4, GRIP2, NUP210 and TLR5 could not be linked to the SR phenomenon by public knowledge.

In summary, this results in a list of **five genes** to be considered as being potentially linked to the SR phenomenon: **C3, CD86, P3H3, STUB1 and KISS1**.

[42] https://www.ncbi.nlm.nih.gov/protein/NP_775960.4

4.4 Conclusions

Skin rash occurrence or absence upon Cetuximab treatment is used as predictive phenotypic biomarker, but still not understood in terms of the underlying molecular mechanisms. For some cases, polymorphisms of EGFR have been described as being relevant, as binding affinity of the therapeutic antibody might be negatively influenced [Jaka *et al.* 2014]. Also, for Fcγ receptor (FcγR) such polymorphisms have been reported and suggested for lowering the efficacy of antibody-dependent cell-mediated cytotoxicity (ADCC) [Bibeau *et al.* 2009]. Both conditions have been precluded for the patients underlying this work.

However, given data with 23 exome datasets was sparse, reasoned i.a. by just two available clinical studies. The Cancer Genome Atlas (TCGA[43]) unfortunately does not offer clinical information on the skin toxicity grade. Hence, statistically a quite low sample size is opposed to a very high number of variables, which in this setting are the genomic variants. Consequently, just an exploratory analysis was possible, lacking validation by design [Bickel *et al.* 2009]. Also the imbalance criterion, assuming bionomial distributions, appears simple, but should be regarded as pragmatic, providing basic statistic grounds for primary selection of variants. In order to compensate, a clustering approach on the gene level was performed, taking into account the biological mechanisms of interest only and therefore acting as a filter. Simultaneously, graphical interactive presentation supports hypothesis generation by field experts, as the functional context is instantly visible. This method, implemented in Cytoscape, has been named 'Molecular Systems Map' (MSM) and is the technical core outcome of this work. Despite some technical shortcomings, the advantages of the MSM method could be demonstrated as filter, visualization and platform for graph algorithms. In contrast to initial considerations, no obvious enrichment of detected gene variants could be detected in domains of the MSM (comp. Fig. 19).

Considering the central biomedical question, in this work five candidates overall could be named, associated with findings and provided with an hypothesis each: C3, CD86, P3H3, STUB1 and KISS1. This indicates them to be considerable for further investigations. According to the depth of the developed hypotheses, C3 and P3H3 (Subsections 4.3.1 and 4.3.8, respectively), appear most worth to be reviewed. For those, the respective mechanistic models should be considered as working hypotheses for future validation studies. These models include proteins being connected to either C3 or P3H3 according to literature and therefore offer several molecular targets to measure. As both models imply balance shifts in regulation, quantitative experiments appear most supportive for determining the background of discriminative skin rash phenotypes upon Cetuximab treatment on a mid-term perspective.

[43] http://cancergenome.nih.gov

For C3 (comp. Subsection 4.3.1), major implications concern immune system mechanisms. The protein is the regulative core element of the complement cascade, which is simultaneously, although accounted for innate immunity, tightly linked to adaptive immune systeme mechanisms. Consequently, a range of immobiochemistry assays might help to quantify the reactivity of C3 and its effects on complement cascade as well as T cell-mediated anti-tumor activity.

For P3H3 (comp. Subsection 4.3.8), also immune system effects should be considered to be measured, but according to the enzyme's proposed influence on collagen type XVII, cell culture-based assays should investigate on this yet hypothetical relation and its mechanistic role in hemidesmosome assembly and disassembly. If upon EGFR inhibition measureable differences occur depending on the P3H3 phenotype, other players of the proposed model (e.g. sheddases) might be interesting to considered as additional targets.

In spite of everything, one has to consider the quite low number of individuals taken into account for this work. Still, these are just eleven patients of the group of interest (SR-negative), with limited quality of the underlying NGS data. Hence, a targeted re-sequencing of the claimed loci in more patients of both groups should be performed first for the sake of experimental validation. Therefore, for those loci a chip layout would have to be designed and applied on additional patients, of which some more are listed in the CIOX and FIRE3 study. In contrast, referencing public datasets like TCGA for validation appears difficult, as like stated above clinical information like survival times and especially the skin rash status are not available.

Coming back to the methodology, one candidate revealed a usability issue in the current implementation of the MSM: while four candidates (C3, CD86, P3H3 and STUB1) have been located in the MSM, KISS1 has been missed; its identification depends on the list of imbalanced candidates only. Although this gene offers an interesting hypothesis, it was simply not contained in any of the broadly selected pathways underlying the skin rash MSM. However, it could be found in the most recent functional interaction list of Reactome.

This observation points out one current drawback of the MSM: while the calling of imbalanced variants is completely automated, the generation of the target map still includes a series of manual steps. Consequently, the MSM particularly developed for this work remained static since the point where variants had been (semi-automatically) mapped to it. A higher degree of automation within the Cytoscape environment is therefore necessary and planned (comp. chapter 5).

Summing up, the MSM principle appears well applicable for closing the gap between high-throughput genomic data (e.g. exome sequencing) and domain experts in the course of explorative analysis. This is primarely due to the biology-based restriction to the relevant contextes, but also to the use of a

graphical representation in an interactive environment. Projecting higher degrees of automation, even for non-computer scientists the use case-specific adaption of the MSM principle appears easy: choosing a formatted pathway database, picking the subsets of interest, remove undesired nodes or edges, and merge all networks to an MSM. As soon as the tasks can be reduced to these major points, ommitting all non-exporatory editing work in text files or the graphical workbench, any domain user is capable of performing hypothesis generation beyond scrolling annotated variant call or gene lists. Importantly, validation studies have to be considered.

5. Outlook

Methodologically, the MSM implementation should be considered for improvements, especially in terms of automation. The example of KISS1 demonstrated that the manual multistep generation of the map from Reactome pathways may lead to update issues, as the integration of new knowledge is comparably cumbersome. Information on functional interactions are provided on an roughly annual basis [Wu *et al.* 2014], and a MSM generated once does not motivate to start the process from scratch when such an update is released. The main reason is the necessity to add attribute columns on origin to both the nodes and edges when iteratively merging the FI networks converted from Reactome pathways. While the process repeated clicking and labeling is not sophisticated at all, it it both error-prone and annoying. In turn, this makes it well suitable for a programmatic solution. Cytoscape comes with a constantly growing application programming interface (API; [Ono *et al.* 2015]), offering the possibility for both automated data analysis and visualization workflows[44] by interconnecting Cytoscape with Python[45], Jupyter notebooks[46] or RStudio[47].

Although the capabilities and limitations remain subject to further evaluation, Python as a programming language and Jupyter notebooks or RStudio as so-called interactive environments predestines the whole approach for integration into Galaxy [Afgan *et al.* 2016]. Galaxy is currently the most widely used workflow and already hosting the genomic variant calling workflow in the given use case. As a data processing platform for various high-throughput data, especially from the 'omics field, also the mapping of other 'omics-derived information should be considered to be mapped to the MSM. Furthermore, selecting subsets of the graph (i.e. genes of interest) may lead to continuative questions on further data, e.g. from other studies or public resources. These could be seemlessly treated in Galaxy, as the integrated interactive environments allow a backport of results. For the sake of reproducibility and documentation, the graph-related analyses can be saved as a notebook, being part of Galaxy workflow like the tools used in common 'omics analyses (comp. Section 2.3). However, such advanced, notebook-based approaches might be primarely dedicated to computational biologists, bioinformaticians or data scientists.

Alternatively, with Cytoscape.js[48] a JavaScript implementation is available, allowing for less encapsulated, more straight-forward visualization of the MSM, suitable for biologists or clinicians. In such settings, graph-related processing steps could be implemented as classical Galaxy workflows of

[44] http://apps.cytoscape.org/apps/cyrest
[45] https://www.python.org
[46] http://jupyter.org
[47] https://www.rstudio.com
[48] http://js.cytoscape.org

regular, configurable tools. In any case, the deeper integration and automation of the MSM principle desires attention, as it also opens the principle to a broader community in a user-friendly way. Correspondingly, it would further reduce the fields expert's dependance on support by computer scientist.

In terms of providing further biological information for hypothesis refinement an integrated connection to literature knowledge appears desirable. In fact, in the course of this work, such information had to be collected by browsing through portals like PubMed, GeneCards, dbSNP and others. This included the generation of queries, checking cross-references and comprehending information in various formats, including tables, figures and and full texts. Although this meets in parts the biologists classical way of dealing with such information and several knowledge sources already list particular evidences instead of just publication references, it appears time consuming. For sure, PubMed for example offers the possibility for restricting queries to findings in human experiments. However, a fully semantic search engine like SCAIview[49] [Younesi *et al.* 2012] would also allow more precise filters like presenting publications with cancers of the gastrointestinal tract or skin, only. Also, cross-reading would be speeded up by class-specific entity highlighting. Modeling such public information again as networks (e.g. using the Biological Expression Language (BEL[50]); comp. Slater 2014) offers chances for faster and deeper integration of knowledge as well as the identification of burried links. Consequently, this could add further support for hypothesis generation, as such networks could be interconnected with the MSM.

Scientifically, the five candidates should be subject to deeper discussion with biomedical experts from oncology, immunology, cell biology and related fields. Targeted re-sequencing should be aimed, starting with the identification of suitable patients offering both biomaterial and clinical information on treatment and adverse effects. If a relevant collective can be assembled, a chip design for the loci of interest would have to be created, enabling deep sequencing e.g. on an Illumina MiSeq platform. From this, the developed workflow could be principally re-used. Anyhow, variant calls from other platform like Ion Torrent PGM/Proton are also applicable, but might require appropriate setup of an adequate workflow.

Generally, the MSM principle is assumed to have the potential to serve as helpful tool in biomedical research for such settings where outcomes of untargeted approaches do not provide hypotheses for further investigations. As the presented principle depends on genes of interest (e.g. those attributed with called variants) and pathways being assembled to a functional interaction network in a usecase-specific manner, it is not restricted to a certain particular field in biomedical research. Currently, an

[49] http://academia.scaiview.com/academia/
[50] http://openbel.org

MSM is being set up for a usecase in neurodegeneration (amyotrophic lateral sclerosis; ALS), where prior analyses revealed a limited number of genes, for which mechanistic hypothesis on the mode of action are lacking. The MSM principle is suggested to bring these genes into a broader functional context. Nevertheless, the unvalidated status of the hypotheses generated in such way have to be taken into account, claiming for further lab work.

6. References

Ab Mutalib, Nurul-Syakima; Md Yusof, Najwa F.; Abdul, Shafina-Nadiawati; Jamal, Rahman (2017) - Pharmacogenomics DNA Biomarkers in Colorectal Cancer: Current Update. In *Frontiers in pharmacology* 8, p. 736. DOI: 10.3389/fphar.2017.00736.

Aboud-Pirak, E.; Hurwitz, E.; Pirak, M. E.; Bellot, F.; Schlessinger, J.; Sela, M. (1988) - Efficacy of Antibodies to Epidermal Growth Factor Receptor Against KB Carcinoma In Vitro and in Nude Mice. In *JNCI Journal of the National Cancer Institute* 80 (20), pp. 1605–1611. DOI: 10.1093/jnci/80.20.1605.

Abu-Humaidan, Anas H. A.; Ananthoju, Nageshwar; Mohanty, Tirthankar; Sonesson, Andreas; Alberius, Per; Schmidtchen, Artur et al. (2014) - The epidermal growth factor receptor is a regulator of epidermal complement component expression and complement activation. In *Journal of immunology (Baltimore, Md. : 1950)* 192 (7), pp. 3355–3364. DOI: 10.4049/jimmunol.1302305.

Afgan, Enis; Baker, Dannon; van den Beek, Marius; Blankenberg, Daniel; Bouvier, Dave; Čech, Martin et al. (2016) - The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. In *Nucleic acids research* 44 (W1), W3-W10. DOI: 10.1093/nar/gkw343.

Aho, Alfred V.; Kernighan, Brian W.; Weinberger, Peter J. (1988) - The AWK programming language. Repr. with corr. Reading, Mass.: Addison-Wesley (Addison-Wesley series in computer science).

Aiba, Yoshihiro; Nakamura, Minoru; Joshita, Satoru; Inamine, Tatsuo; Komori, Atsumasa; Yoshizawa, Kaname et al. (2011) - Genetic polymorphisms in CTLA4 and SLC4A2 are differentially associated with the pathogenesis of primary biliary cirrhosis in Japanese patients. In *Journal of gastroenterology* 46 (10), pp. 1203–1212. DOI: 10.1007/s00535-011-0417-7.

Antonov, Dimitar; Kazandjieva, Jana; Etugov, Doncho; Gospodinov, Dimitar; Tsankov, Nikolai (2004) - Drug-induced lupus erythematosus. In *Clinics in dermatology* 22 (2), pp. 157–166. DOI: 10.1016/j.clindermatol.2003.12.023.

Arribas, Joaquín; Esselens, Cary (2009) - ADAM17 as a therapeutic target in multiple diseases. In *Current pharmaceutical design* 15 (20), pp. 2319–2335.

Auffray, Charles; Balling, Rudi; Barroso, Inês; Bencze, László; Benson, Mikael; Bergeron, Jay et al. (2016) - Making sense of big data in health research: Towards an EU action plan. In *Genome medicine* 8 (1), p. 71. DOI: 10.1186/s13073-016-0323-y.

Auton, Adam; Brooks, Lisa D.; Durbin, Richard M.; Garrison, Erik P.; Kang, Hyun Min; Korbel, Jan O. et al. (2015) - A global reference for human genetic variation. In *Nature* 526 (7571), pp. 68–74. DOI: 10.1038/nature15393.

Azeloglu, Evren U.; Iyengar, Ravi (2015) - Signaling networks: information flow, computation, and decision making. In *Cold Spring Harbor perspectives in biology* 7 (4), a005934. DOI: 10.1101/cshperspect.a005934.

Bager, C. L.; Karsdal, M. A. (2016) - Type XVIII Collagen. In Morten Karsdal (Ed.): Biochemistry of collagens. Structure, function and biomarkers. London, United Kingdom: Academic Press, pp. 113–121.

Bartkowiak, Bartlomiej; Greenleaf, Arno L. (2011) - Phosphorylation of RNAPII: To P-TEFb or not to P-TEFb? In *Transcription* 2 (3), pp. 115–119. DOI: 10.4161/trns.2.3.15004.

Beck, Benjamin H.; Welch, Danny R. (2010) - The KISS1 metastasis suppressor: A good night kiss for disseminated cancer cells. In *European journal of cancer (Oxford, England : 1990)* 46 (7), pp. 1283–1289. DOI: 10.1016/j.ejca.2010.02.023.

Berglund, Nils A.; Kargas, Vasileios; Ortiz-Suarez, Maite L.; Bond, Peter J. (2015) - The role of protein-protein interactions in Toll-like receptor function. In *Progress in biophysics and molecular biology* 119 (1), pp. 72–83. DOI: 10.1016/j.pbiomolbio.2015.06.021.

Bergstrom, Carl T.; Antia, Rustom (2006) - How do adaptive immune systems control pathogens while avoiding autoimmunity? In *Trends in ecology & evolution* 21 (1), pp. 22–28. DOI: 10.1016/j.tree.2005.11.008.

Berman, Jules J. (2004) - Tumor classification: Molecular analysis meets Aristotle. In *BMC cancer* 4, p. 10. DOI: 10.1186/1471-2407-4-10.

Bhavnani, Suresh K.; Drake, Justin; Divekar, Rohit (2014) - The role of visual analytics in asthma phenotyping and biomarker discovery. In *Advances in experimental medicine and biology* 795, pp. 289–305. DOI: 10.1007/978-1-4614-8603-9_18.

Bibeau, Frédéric; Lopez-Crapez, Evelyne; Di Fiore, Frédéric; Thezenas, Simon; Ychou, Marc; Blanchard, France et al. (2009) - Impact of Fc{gamma}RIIa-Fc{gamma}RIIIa polymorphisms and KRAS mutations on the clinical outcome of patients with metastatic colorectal cancer treated with cetuximab plus irinotecan. In *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 27 (7), pp. 1122–1129. DOI: 10.1200/JCO.2008.18.0463.

Bickel, Peter J.; Brown, James B.; Huang, Haiyan; Li, Qunhua (2009) - An overview of recent developments in genomics and associated statistical methods. In *Philosophical transactions. Series*

A, Mathematical, physical, and engineering sciences 367 (1906), pp. 4313–4337. DOI: 10.1098/rsta.2009.0164.

Bignucolo, Alessia; Mattia, Elena de; Cecchin, Erika; Roncato, Rossana; Toffoli, Giuseppe (2017) - Pharmacogenomics of Targeted Agents for Personalization of Colorectal Cancer Treatment. In *International journal of molecular sciences* 18 (7). DOI: 10.3390/ijms18071522.

Blake, Alexandra; Dragan, Magdalena; Tirona, Rommel G.; Hardy, Daniel B.; Brackstone, Muriel; Tuck, Alan B. et al. (2017) - G protein-coupled KISS1 receptor is overexpressed in triple negative breast cancer and promotes drug resistance. In *Scientific reports* 7, p. 46525. DOI: 10.1038/srep46525.

Blazek, Dalibor; Kohoutek, Jiri; Bartholomeeusen, Koen; Johansen, Eric; Hulinkova, Petra; Luo, Zeping et al. (2011) - The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. In *Genes & development* 25 (20), pp. 2158–2172. DOI: 10.1101/gad.16962311.

Bork, P.; Holm, L.; Sander, C. (1994) - The immunoglobulin fold. Structural classification, sequence patterns and common core. In *Journal of molecular biology* 242 (4), pp. 309–320. DOI: 10.1006/jmbi.1994.1582.

Bösken, Christian A.; Farnung, Lucas; Hintermair, Corinna; Merzel Schachter, Miriam; Vogel-Bachmayr, Karin; Blazek, Dalibor et al. (2014) - The structure and substrate specificity of human Cdk12/Cyclin K. In *Nature communications* 5, p. 3505. DOI: 10.1038/ncomms4505.

Brackett, Craig M.; Kojouharov, Bojidar; Veith, Jean; Greene, Kellee F.; Burdelya, Lyudmila G.; Gollnick, Sandra O. et al. (2016) - Toll-like receptor-5 agonist, entolimod, suppresses metastasis and induces immunity by stimulating an NK-dendritic-CD8+ T-cell axis. In *Proceedings of the National Academy of Sciences of the United States of America* 113 (7), E874-83. DOI: 10.1073/pnas.1521359113.

Bruckner-Tuderman, Leena; Has, Cristina (2012) - Molecular heterogeneity of blistering disorders: The paradigm of epidermolysis bullosa. In *The Journal of investigative dermatology* 132 (E1), E2-5. DOI: 10.1038/skinbio.2012.2.

Brunner, M. C.; Chambers, C. A.; Chan, F. K.; Hanke, J.; Winoto, A.; Allison, J. P. (1999) - CTLA-4-Mediated inhibition of early events of T cell proliferation. In *Journal of immunology (Baltimore, Md. : 1950)* 162 (10), pp. 5813–5820.

Bujko, Mateusz; Kober, Paulina; Mikula, Michal; Ligaj, Marcin; Ostrowski, Jerzy; Siedlecki, Janusz Aleksander (2015) - Expression changes of cell-cell adhesion-related genes in colorectal tumors. In *Oncology letters* 9 (6), pp. 2463–2470. DOI: 10.3892/ol.2015.3107.

Candi, Eleonora; Schmidt, Rainer; Melino, Gerry (2005) - The cornified envelope: A model of cell death in the skin. In *Nature reviews. Molecular cell biology* 6 (4), pp. 328–340. DOI: 10.1038/nrm1619.

Carmona, F. Javier; Villanueva, Alberto; Vidal, August; Muñoz, Clara; Puertas, Sara; Penin, Rosa M. et al. (2012) - Epigenetic disruption of cadherin-11 in human cancer metastasis. In *The Journal of pathology* 228 (2), pp. 230–240. DOI: 10.1002/path.4011.

Carpenter, Graham; King, Lloyd; Cohen, Stanley (1978) - Epidermal growth factor stimulates phosphorylation in membrane preparations in vitro. In *Nature* 276 (5686), pp. 409–410. DOI: 10.1038/276409a0.

Cescon, Matilde; Gattazzo, Francesca; Chen, Peiwen; Bonaldo, Paolo (2015) - Collagen VI at a glance. In *Journal of cell science* 128 (19), pp. 3525–3531. DOI: 10.1242/jcs.169748.

Chen, Peiwen; Cescon, Matilde; Bonaldo, Paolo (2013a) - Collagen VI in cancer and its biological mechanisms. In *Trends in molecular medicine* 19 (7), pp. 410–417. DOI: 10.1016/j.molmed.2013.04.001.

Chen, Zuojia; Barbi, Joseph; Bu, Shurui; Yang, Huang-Yu; Li, Zhiyuan; Gao, Yayi et al. (2013b) - The ubiquitin ligase Stub1 negatively modulates regulatory T cell suppressive activity by promoting degradation of the transcription factor Foxp3. In *Immunity* 39 (2), pp. 272–285. DOI: 10.1016/j.immuni.2013.08.006.

Cho, Hyun-Soo; Leahy, Daniel J. (2002) - Structure of the extracellular region of HER3 reveals an interdomain tether. In *Science (New York, N.Y.)* 297 (5585), pp. 1330–1333. DOI: 10.1126/science.1074611.

Chowdhury, S. J.; Karra, V. K.; Gumma, P. K.; Bharali, R.; Kar, P. (2015) - rs2230201 polymorphism may dictate complement C3 levels and response to treatment in chronic hepatitis C patients. In *Journal of viral hepatitis* 22 (2), pp. 184–191. DOI: 10.1111/jvh.12280.

Chung, Chaeuk; Yoo, Geon; Kim, Tackhoon; Lee, Dahye; Lee, Choong-Sik; Cha, Hye Rim et al. (2016) - The E3 ubiquitin ligase CHIP selectively regulates mutant epidermal growth factor receptor by ubiquitination and degradation. In *Biochemical and biophysical research communications* 479 (2), pp. 152–158. DOI: 10.1016/j.bbrc.2016.07.111.

Chung, Ki Young; Shia, Jinru; Kemeny, Nancy E.; Shah, Manish; Schwartz, Gary K.; Tse, Archie et al. (2005) - Cetuximab shows activity in colorectal cancer patients with tumors that do not express the epidermal growth factor receptor by immunohistochemistry. In *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 23 (9), pp. 1803–1810. DOI: 10.1200/JCO.2005.08.037.

Ciliberto, Domenico; Staropoli, Nicoletta; Caglioti, Francesca; Chiellino, Silvia; Ierardi, Antonella; Ingargiola, Rossana et al. (2018) - The best strategy for RAS wild-type metastatic colorectal cancer patients in first-line treatment: A classic and Bayesian meta-analysis. In *Critical reviews in oncology/hematology* 125, pp. 69–77. DOI: 10.1016/j.critrevonc.2018.03.003.

Cingolani, Pablo; Platts, Adrian; Le Wang, Lily; Coon, Melissa; Nguyen, Tung; Wang, Luan et al. (2012) - A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. In *Fly* 6 (2), pp. 80–92. DOI: 10.4161/fly.19695.

Cipollini, Monica; Landi, Stefano; Gemignani, Federica (2014) - MicroRNA binding site polymorphisms as biomarkers in cancer management and research. In *Pharmacogenomics and personalized medicine* 7, pp. 173–191. DOI: 10.2147/PGPM.S61693.

Cohen, Roger B. (2003) - Epidermal growth factor receptor as a therapeutic target in colorectal cancer. In *Clinical colorectal cancer* 2 (4), pp. 246–251. DOI: 10.3816/CCC.2003.n.006.

Croft, David; Mundo, Antonio Fabregat; Haw, Robin; Milacic, Marija; Weiser, Joel; Wu, Guanming et al. (2014) - The Reactome pathway knowledgebase. In *Nucleic acids research* 42 (Database issue), D472-7. DOI: 10.1093/nar/gkt1102.

da Silva, Karina Ribeiro; Fraga, Tatiana Rodrigues; Lucatelli, Juliana Faggion; Grumach, Anete Sevciovic; Isaac, Lourdes (2016) - Skipping of exon 27 in C3 gene compromises TED domain and results in complete human C3 deficiency. In *Immunobiology* 221 (5), pp. 641–649. DOI: 10.1016/j.imbio.2016.01.005.

Danecek, Petr; Auton, Adam; Abecasis, Goncalo; Albers, Cornelis A.; Banks, Eric; DePristo, Mark A. et al. (2011) - The variant call format and VCFtools. In *Bioinformatics (Oxford, England)* 27 (15), pp. 2156–2158. DOI: 10.1093/bioinformatics/btr330.

Danese, Silvio; Malesci, Alberto; Vetrano, Stefania (2011) - Colitis-associated cancer: The dark side of inflammatory bowel disease. In *Gut* 60 (12), pp. 1609–1610. DOI: 10.1136/gutjnl-2011-300953.

Dastani, Mehdi (2002) - The Role of Visual Perception in Data Visualization. In *Journal of Visual Languages & Computing* 13 (6), pp. 601–622. DOI: 10.1006/jvlc.2002.0235.

Deschoolmeester, Vanessa; Baay, Marc; Specenier, Pol; Lardon, Filip; Vermorken, Jan B. (2010) - A review of the most promising biomarkers in colorectal cancer: One step closer to targeted therapy. In *The oncologist* 15 (7), pp. 699–731. DOI: 10.1634/theoncologist.2010-0025.

Dika, Emi; Ravaioli, Giulia Maria; Fanti, Pier Alessandro; Piraccini, Bianca Maria; Lambertini, Martina; Chessa, Marco Adriano et al. (2017) - Cutaneous adverse effects during ipilimumab treatment for

metastatic melanoma: a prospective study. In *European journal of dermatology : EJD* 27 (3), pp. 266–270. DOI: 10.1684/ejd.2017.3023.

Duke, Susan P.; Bancken, Fabrice; Crowe, Brenda; Soukup, Mat; Botsis, Taxiarchis; Forshee, Richard (2015) - Seeing is believing: Good graphic design principles for medical research. In *Statistics in medicine* 34 (22), pp. 3040–3059. DOI: 10.1002/sim.6549.

Eberl, Gérard (2016) - Immunity by equilibrium. In *Nature reviews. Immunology* 16 (8), pp. 524–532. DOI: 10.1038/nri.2016.75.

Eilers, R. E.; Gandhi, M.; Patel, J. D.; Mulcahy, M. F.; Agulnik, M.; Hensing, T.; Lacouture, Mario E. (2010) - Dermatologic infections in cancer patients treated with epidermal growth factor receptor inhibitor therapy. In *Journal of the National Cancer Institute* 102 (1), pp. 47–53. DOI: 10.1093/jnci/djp439.

Ekumi, Kingsley M.; Paculova, Hana; Lenasi, Tina; Pospichalova, Vendula; Bösken, Christian A.; Rybarikova, Jana et al. (2015) - Ovarian carcinoma CDK12 mutations misregulate expression of DNA repair genes via deficient formation and function of the Cdk12/CycK complex. In *Nucleic acids research* 43 (5), pp. 2575–2589. DOI: 10.1093/nar/gkv101.

Elmgreen, Jens; Sørensen, Henning; Berkowicz, Adela (1984) - Polymorphism of Complement C3 in Chronic Inflammatory Bowel Disease. In *Acta Medica Scandinavica* 215 (4), pp. 375–378. DOI: 10.1111/j.0954-6820.1984.tb05021.x.

EMA (2009) - European public assessment report (EPAR). European Medicines Agency. Available online at http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/human/medicines/001016/human_med_000857.jsp&mid=WC0b01ac058001d124, updated on 11/11/2009.

evaluate.com (2002) - AstraZeneca Secures First Market Approval For Iressa™ In Japan. Available online at http://www.evaluategroup.com/Universal/View.aspx?type=Story&id=28010, updated on 7/5/2002.

Ferlay, Jacques; Shin, Hai-Rim; Bray, Freddie; Forman, David; Mathers, Colin; Parkin, Donald Maxwell (2010) - Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. In *International journal of cancer* 127 (12), pp. 2893–2917. DOI: 10.1002/ijc.25516.

Few, Stephen (2013) - Data Visualization for Human Perception: Chapter 35. In Mads Soegaard, Rikke Friis Dam (Eds.): The Encyclopedia of Human-Computer Interaction. 2nd ed. Aarhus, Denmark: The Interaction Design Foundation.

Fitts, Braxton; Zhang, Ziran; Maher, Massoud; Demchak, Barry (2016) - dot-app: A Graphviz-Cytoscape conversion plug-in. In *F1000Research* 5, p. 2543. DOI: 10.12688/f1000research.9751.2.

Fraser-Pitt, Douglas J.; Cameron, Pamela; McNeilly, Tom N.; Boyd, Amanda; Manson, Erin D. T.; Smith, David G. E. (2011) - Phosphorylation of the epidermal growth factor receptor (EGFR) is essential for interleukin-8 release from intestinal epithelial cells in response to challenge with Escherichia coli O157: H7 flagellin. In *Microbiology (Reading, England)* 157 (Pt 8), pp. 2339–2347. DOI: 10.1099/mic.0.047670-0.

Fruchterman, Thomas M. J.; Reingold, Edward M. (1991) - Graph drawing by force-directed placement. In *Softw: Pract. Exper.* 21 (11), pp. 1129–1164. DOI: 10.1002/spe.4380211102.

Fuchs, Elaine; Raghavan, Srikala (2002) - Getting under the skin of epidermal morphogenesis. In *Nature reviews. Genetics* 3 (3), pp. 199–209. DOI: 10.1038/nrg758.

Gan, Lei; Liu, Dong-Bo; Lu, Hai-Feng; Long, Guo-Xian; Mei, Qi; Hu, Guang-Yuan et al. (2012) - Decreased expression of the carboxyl terminus of heat shock cognate 70 interacting protein in human gastric cancer and its clinical significance. In *Oncology reports* 28 (4), pp. 1392–1398. DOI: 10.3892/or.2012.1957.

Gao, Nan; Kumar, Ashok; Jyot, Jeevan; Yu, Fu-Shin (2010) - Flagellin-induced corneal antimicrobial peptide production and wound repair involve a novel NF-kappaB-independent and EGFR-dependent pathway. In *PloS one* 5 (2), e9351. DOI: 10.1371/journal.pone.0009351.

Gara, Sudheer Kumar; Grumati, Paolo; Squarzoni, Stefano; Sabatelli, Patrizia; Urciuolo, Anna; Bonaldo, Paolo et al. (2011) - Differential and restricted expression of novel collagen VI chains in mouse. In *Matrix biology : journal of the International Society for Matrix Biology* 30 (4), pp. 248–257. DOI: 10.1016/j.matbio.2011.03.006.

Gaude, H.; Aznar, N.; Delay, A.; Bres, A.; Buchet-Poyau, K.; Caillat, C. et al. (2012) - Molecular chaperone complexes with antagonizing activities regulate stability and activity of the tumor suppressor LKB1. In *Oncogene* 31 (12), pp. 1582–1591. DOI: 10.1038/onc.2011.342.

Gazdar, Adi F. (2009) - Personalized medicine and inhibition of EGFR signaling in lung cancer. In *The New England journal of medicine* 361 (10), pp. 1018–1020. DOI: 10.1056/NEJMe0905763.

Gerogianni, Kalliopi; Tsezou, Aspasia; Dimas, Konstantinos (2018) - Drug-Induced Skin Adverse Reactions: The Role of Pharmacogenomics in Their Prevention. In *Molecular diagnosis & therapy. DOI:* 10.1007/s40291-018-0330-3.

Giebeler, Nives; Zigrino, Paola (2016) - A Disintegrin and Metalloprotease (ADAM): Historical Overview of Their Functions. In *Toxins* 8 (4), p. 122. DOI: 10.3390/toxins8040122.

Giudice, G. J.; Emery, D. J.; Diaz, L. A. (1992) - Cloning and primary structural analysis of the bullous pemphigoid autoantigen BP180. In *The Journal of investigative dermatology* 99 (3), pp. 243–250.

Gjaltema, Rutger A. F.; Bank, Ruud A. (2017) - Molecular insights into prolyl and lysyl hydroxylation of fibrillar collagens in health and disease. In *Critical reviews in biochemistry and molecular biology* 52 (1), pp. 74–95. DOI: 10.1080/10409238.2016.1269716.

Gooz, Monika (2010) - ADAM-17: The enzyme that does it all. In *Critical reviews in biochemistry and molecular biology* 45 (2), pp. 146–169. DOI: 10.3109/10409231003628015.

Grandis, Jennifer Rubin; Sok, John C. (2004) - Signaling through the epidermal growth factor receptor during the development of malignancy. In *Pharmacology & therapeutics* 102 (1), pp. 37–46. DOI: 10.1016/j.pharmthera.2004.01.002.

Gray, Stacy W.; Gagan, Jeffrey; Cerami, Ethan; Cronin, Angel M.; Uno, Hajime; Oliver, Nelly et al. (2018) - Interactive or static reports to guide clinical interpretation of cancer genomics. In *Journal of the American Medical Informatics Association : JAMIA. DOI:* 10.1093/jamia/ocx150.

Greber, U. F.; Senior, A.; Gerace, L. (1990) - A major glycoprotein of the nuclear pore complex is a membrane-spanning polypeptide with a large lumenal domain and a small cytoplasmic tail. In *The EMBO Journal* 9 (5), pp. 1495–1502.

Gruenwald, Katrin; Castagnola, Patrizio; Besio, Roberta; Dimori, Milena; Chen, Yuqing; Akel, Nisreen S. et al. (2014) - Sc65 is a novel endoplasmic reticulum protein that regulates bone mass homeostasis. In *Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research* 29 (3), pp. 666–675. DOI: 10.1002/jbmr.2075.

Gude, Dilip (2011) - Hair follicle stem cells: A new arena. In *International journal of trichology* 3 (2), pp. 125–126. DOI: 10.4103/0974-7753.90840.

Guo, Yu; Zhao, Ming; Lu, Qianjin (2016) - Transcription factor RFX1 is ubiquitinated by E3 ligase STUB1 in systemic lupus erythematosus. In *Clinical immunology (Orlando, Fla.)* 169, pp. 1–7. DOI: 10.1016/j.clim.2016.06.003.

Hanahan, Douglas; Weinberg, Robert A. (2000) - The Hallmarks of Cancer. In *Cell* 100 (1), pp. 57–70. DOI: 10.1016/S0092-8674(00)81683-9.

Hanahan, Douglas; Weinberg, Robert A. (2011) - Hallmarks of cancer: The next generation. In *Cell* 144 (5), pp. 646–674. DOI: 10.1016/j.cell.2011.02.013.

Harms, John F.; Welch, Danny R. (2002) - The Role of KISS1 in Melanoma Metastasis Suppression. In Danny R. Welch (Ed.): Cancer metastasis, related genes. Dordrecht: Kluwer Academic (Cancer metastasis, v. 3), pp. 219–229.

Harris, Claire L.; Heurich, Meike; Rodriguez de Cordoba, Santiago; Morgan, B. Paul (2012) - The complotype: Dictating risk for inflammation and infection. In *Trends in immunology* 33 (10), pp. 513–521. DOI: 10.1016/j.it.2012.06.001.

Hatzimichael, E.; Lo Nigro, C.; Lattanzio, L.; Syed, N.; Shah, R.; Dasoula, A. et al. (2012) - The collagen prolyl hydroxylases are novel transcriptionally silenced genes in lymphoma. In *British journal of cancer* 107 (8), pp. 1423–1432. DOI: 10.1038/bjc.2012.380.

Hawn, Thomas R.; Wu, Hui; Grossman, Jennifer M.; Hahn, Bevra H.; Tsao, Betty P.; Aderem, Alan (2005) - A stop codon polymorphism of Toll-like receptor 5 is associated with resistance to systemic lupus erythematosus. In *Proceedings of the National Academy of Sciences of the United States of America* 102 (30), pp. 10593–10597. DOI: 10.1073/pnas.0501165102.

Heard, Melissa E.; Besio, Roberta; Weis, MaryAnn; Rai, Jyoti; Hudson, David M.; Dimori, Milena et al. (2016) - Sc65-Null Mice Provide Evidence for a Novel Endoplasmic Reticulum Complex Regulating Collagen Lysyl Hydroxylation. In *PLoS genetics* 12 (4), e1006002. DOI: 10.1371/journal.pgen.1006002.

Heimdal, Ketil; Sanchez-Guixé, Monica; Aukrust, Ingvild; Bollerslev, Jens; Bruland, Ove; Jablonski, Greg Eigner et al. (2014) - STUB1 mutations in autosomal recessive ataxias - evidence for mutation-specific clinical heterogeneity. In *Orphanet journal of rare diseases* 9, p. 146. DOI: 10.1186/s13023-014-0146-0.

Heinemann, Volker; Weikersthal, Ludwig Fischer von; Decker, Thomas; Kiani, Alexander; Vehling-Kaiser, Ursula; Al-Batran, Salah-Eddin et al. (2014) - FOLFIRI plus cetuximab versus FOLFIRI plus bevacizumab as first-line treatment for patients with metastatic colorectal cancer (FIRE-3): A randomised, open-label, phase 3 trial. In *The Lancet Oncology* 15 (10), pp. 1065–1075. DOI: 10.1016/S1470-2045(14)70330-4.

Helmholtz, Hermann von (Ed.) (1867) - Handbuch der physiologischen Optik. 3 volumes. Leipzig: Voss (Handbuch der physiologischen Optik, 3).

Herbst, Roy S. (2003) - With Cetuximab and Erlotinib, Rash Correlates With Survival. In *Oncology* 12 (11). Available online at http://www.cancernetwork.com/articles/cetuximab-and-erlotinib-rash-correlates-survival.

Herbst, Roy S.; LoRusso, Patricia M.; Purdom, Michele; Ward, Deborah (2003) - Dermatologic Side Effects Associated with Gefitinib Therapy: Clinical Experience and Management. In *Clinical Lung Cancer* 4 (6), pp. 366–369. DOI: 10.3816/CLC.2003.n.016.

Hesling, C.; D'Incan, M.; Mansard, S.; Franck, F.; Corbin-Duval, A.; Chèvenet, C. et al. (2004) - In vivo and in situ modulation of the expression of genes involved in metastasis and angiogenesis in a patient treated with topical imiquimod for melanoma skin metastases. In *The British journal of dermatology* 150 (4), pp. 761–767. DOI: 10.1111/j.0007-0963.2004.05898.x.

Hirako, Yoshiaki; Nishizawa, Yuji; Sitaru, Cassian; Opitz, Annika; Marcus, Katrin; Meyer, Helmut E. et al. (2003) - The 97-kDa (LABD97) and 120-kDa (LAD-1) fragments of bullous pemphigoid antigen 180/type XVII collagen have different N-termini. In *The Journal of investigative dermatology* 121 (6), pp. 1554–1556. DOI: 10.1046/j.1523-1747.2003.12607.x.

Holcmann, Martin; Sibilia, Maria (2015) - Mechanisms underlying skin disorders induced by EGFR inhibitors. In *Molecular & cellular oncology* 2 (4), e1004969. DOI: 10.1080/23723556.2015.1004969.

Holland, James F.; Frei, Emil; Kufe, Donald W. (Eds.) (2003) - Cancer medicine: Chapter 5 - Growth Factors and Signal Transduction in Cancer. American Cancer Society. Hamilton, Ont.: Decker.

Holubec, Lubos; Polivka, Jiri; Safanda, Martin; Karas, Michal; Liska, Vaclav (2016) - The Role of Cetuximab in the Induction of Anticancer Immune Response in Colorectal Cancer Treatment. In *Anticancer research* 36 (9), pp. 4421–4426. DOI: 10.21873/anticanres.10985.

Hsu, Yi-Fan; Ajona, Daniel; Corrales, Leticia; Lopez-Picazo, Jose M.; Gurpide, Alfonso; Montuenga, Luis M.; Pio, Ruben (2010) - Complement activation mediates cetuximab inhibition of non-small cell lung cancer tumor growth in vivo. In *Molecular cancer* 9, p. 139. DOI: 10.1186/1476-4598-9-139.

Huang, Sijia; Chaudhary, Kumardeep; Garmire, Lana X. (2017) - More Is Better: Recent Progress in Multi-Omics Data Integration Methods. In *Frontiers in genetics* 8, p. 84. DOI: 10.3389/fgene.2017.00084.

Hudson, David M.; Eyre, David R. (2013) - Collagen prolyl 3-hydroxylation: A major role for a minor post-translational modification? In *Connective tissue research* 54 (4-5), pp. 245–251. DOI: 10.3109/03008207.2013.800867.

Hudson, David M.; Joeng, Kyu Sang; Werther, Rachel; Rajagopal, Abbhirami; Weis, MaryAnn; Lee, Brendan H.; Eyre, David R. (2015) - Post-translationally abnormal collagens of prolyl 3-hydroxylase-2 null mice offer a pathobiological mechanism for the high myopia linked to human LEPREL1 mutations. In *The Journal of biological chemistry* 290 (13), pp. 8613–8622. DOI: 10.1074/jbc.M114.634915.

Hudson, David M.; Weis, MaryAnn; Rai, Jyoti; Joeng, Kyu Sang; Dimori, Milena; Lee, Brendan H. et al. (2017) - P3h3-null and Sc65-null Mice Phenocopy the Collagen Lysine Under-hydroxylation and

Cross-linking Abnormality of Ehlers-Danlos Syndrome Type VIA. In *The Journal of biological chemistry* 292 (9), pp. 3877–3887. DOI: 10.1074/jbc.M116.762245.

Hudson, David M.; Werther, Rachel; Weis, MaryAnn; Wu, Jiann-Jiu; Eyre, David R. (2014) - Evolutionary origins of C-terminal (GPP)n 3-hydroxyproline formation in vertebrate tendon collagen. In *PloS one* 9 (4), e93467. DOI: 10.1371/journal.pone.0093467.

Huo, Xinkai; Zhang, Lei; Li, Tao (2018) - Analysis of the association of the expression of KiSS-1 in colorectal cancer tissues with the pathology and prognosis. In *Oncology letters* 15 (3), pp. 3056–3060. DOI: 10.3892/ol.2017.7630.

Ikeda, Koei; Iyama, Ken-ichi; Ishikawa, Nobuyuki; Egami, Hiroshi; Nakao, Mitsuyoshi; Sado, Yoshikazu et al. (2006) - Loss of Expression of Type IV Collagen α5 and α6 Chains in Colorectal Cancer Associated with the Hypermethylation of Their Promoter Region. In *The American Journal of Pathology* 168 (3), pp. 856–865. DOI: 10.2353/ajpath.2006.050384.

Ishikawa, Yoshihiro; Wirz, Jackie; Vranka, Janice A.; Nagata, Kazuhiro; Bächinger, Hans Peter (2009) - Biochemical characterization of the prolyl 3-hydroxylase 1.cartilage-associated protein.cyclophilin B complex. In *The Journal of biological chemistry* 284 (26), pp. 17641–17647. DOI: 10.1074/jbc.M109.007070.

Itoh, Yusuke; Joh, Takashi; Tanida, Satoshi; Sasaki, Makoto; Kataoka, Hiromi; Itoh, Keisuke et al. (2005) - IL-8 promotes cell proliferation and migration through metalloproteinase-cleavage proHB-EGF in human colon carcinoma cells. In *Cytokine* 29 (6), pp. 275–282. DOI: 10.1016/j.cyto.2004.11.005.

Jaka, Ane; Gutiérrez-Rivera, Araika; Ormaechea, Nerea; Blanco, Jesus; La Casta, Adelaida; Sarasqueta, Cristina et al. (2014) - Association between EGFR gene polymorphisms, skin rash and response to anti-EGFR therapy in metastatic colorectal cancer patients. In *Experimental dermatology* 23 (10), pp. 751–753. DOI: 10.1111/exd.12510.

Janeway, Charles A. (2001) - Immunobiology: The immune system in health and disease ; [animated CD-ROM inside]. 5. ed. New York, NY: Garland Publ. Available online at http://www.ncbi.nlm.nih.gov:80/books/bv.fcgi?call=bv.View.ShowSection&rid=imm.

Joshi, Poorval M.; Sutor, Shari L.; Huntoon, Catherine J.; Karnitz, Larry M. (2014) - Ovarian cancer-associated mutations disable catalytic activity of CDK12, a kinase that promotes homologous recombination repair and resistance to cisplatin and poly(ADP-ribose) polymerase inhibitors. In *The Journal of biological chemistry* 289 (13), pp. 9247–9253. DOI: 10.1074/jbc.M114.551143.

Joshi, Vibhuti; Amanullah, Ayeman; Upadhyay, Arun; Mishra, Ribhav; Kumar, Amit; Mishra, Amit (2016) - A Decade of Boon or Burden: What Has the CHIP Ever Done for Cellular Protein Quality Control

Mechanism Implicated in Neurodegeneration and Aging? In *Frontiers in molecular neuroscience* 9, p. 93. DOI: 10.3389/fnmol.2016.00093.

Jost, M.; Kari, C.; Rodeck, U. (2000) - The EGF receptor - an essential regulator of multiple epidermal functions. In *European journal of dermatology : EJD* 10 (7), pp. 505–510.

Kalluri, Raghu (2003) - Basement membranes: Structure, assembly and role in tumour angiogenesis. In *Nature reviews. Cancer* 3 (6), pp. 422–433. DOI: 10.1038/nrc1094.

Kang, Mingsong; Martin, Alberto (2017) - Microbiome and colorectal cancer: Unraveling host-microbiota interactions in colitis-associated colorectal cancer development. In *Seminars in immunology. DOI:* 10.1016/j.smim.2017.04.003.

Karapetis, Christos S.; Khambata-Ford, Shirin; Jonker, Derek J.; O'Callaghan, Chris J.; Tu, Dongsheng; Tebbutt, Niall C. et al. (2008) - K-ras mutations and benefit from cetuximab in advanced colorectal cancer. In *The New England journal of medicine* 359 (17), pp. 1757–1765. DOI: 10.1056/NEJMoa0804385.

Kato, Kosuke; Lillehoj, Erik P.; Kim, Kwang Chul (2016) - Pseudomonas aeruginosa stimulates tyrosine phosphorylation of and TLR5 association with the MUC1 cytoplasmic tail through EGFR activation. In *Inflammation research : official journal of the European Histamine Research Society … [et al.]* 65 (3), pp. 225–233. DOI: 10.1007/s00011-015-0908-8.

Kato, Kosuke; Lillehoj, Erik P.; Park, Yong Sung; Umehara, Tsuyoshi; Hoffman, Nicholas E.; Madesh, Muniswamy; Kim, K. Chul (2012) - Membrane-tethered MUC1 mucin is phosphorylated by epidermal growth factor receptor in airway epithelial cells and associates with TLR5 to inhibit recruitment of MyD88. In *Journal of immunology (Baltimore, Md. : 1950)* 188 (4), pp. 2014–2022. DOI: 10.4049/jimmunol.1102405.

Khan, M. Afzal; Assiri, A. M.; Broering, D. C. (2015) - Complement and macrophage crosstalk during process of angiogenesis in tumor progression. In *Journal of biomedical science* 22, p. 58. DOI: 10.1186/s12929-015-0151-1.

Kida, Miyuki; Fujioka, Hirotaka; Kosaka, Yoshiyuki; Hayashi, Kouhei; Sakiyama, Yukio; Ariga, Tadashi (2008) - The first confirmed case with C3 deficiency caused by compound heterozygous mutations in the C3 gene; a new aspect of pathogenesis for C3 deficiency. In *Blood cells, molecules & diseases* 40 (3), pp. 410–413. DOI: 10.1016/j.bcmd.2007.11.002.

Kikutake, Chie; Yahara, Koji (2016) - Identification of Epigenetic Biomarkers of Lung Adenocarcinoma through Multi-Omics Data Analysis. In *PloS one* 11 (4), e0152918. DOI: 10.1371/journal.pone.0152918.

Kitajima, Yasuo; Owaribe, Katsushi; Nishizawa, Yuji; Yaoita, Hideo (1992) - Control of the Distribution of Hemidesmosome Components in Cultured Keratinocytes: Ca 2+ and Phorbol Esters. In *The Journal of dermatology* 19 (11), pp. 770–773. DOI: 10.1111/j.1346-8138.1992.tb03778.x.

Kitts, Adrienne; Phan, Lon; Ward, Minghong; Holmes, John Bradley (2013) - The Database of Short Genetic Variation (dbSNP). In Chris Maloney, Ed Sequeira, Christopher Kelly, Rebecca Orris, Jeffrey Beck (Eds.): The NCBI Handbook. 2nd edition, checked on 3/19/2018.

Klahan, Sukhontip; Huang, Chi-Cheng; Chien, Shu-Chen; Wu, Mei-Shin; Wong, Henry Sung-Ching; Huang, Chien-Yu et al. (2016) - Bioinformatic analyses revealed underlying biological functions correlated with oxaliplatin responsiveness. In *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine* 37 (1), pp. 583–590. DOI: 10.1007/s13277-015-3807-2.

Koboldt, Daniel C.; Zhang, Qunyuan; Larson, David E.; Shen, Dong; McLellan, Michael D.; Lin, Ling et al. (2012) - VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. In *Genome research* 22 (3), pp. 568–576. DOI: 10.1101/gr.129684.111.

Kohoutek, Jiri; Blazek, Dalibor (2012) - Cyclin K goes with Cdk12 and Cdk13. In *Cell division* 7, p. 12. DOI: 10.1186/1747-1028-7-12.

Kojouharov, Bojidar M.; Brackett, Craig M.; Veith, Jean M.; Johnson, Christopher P.; Gitlin, Ilya I.; Toshkov, Ilia A. et al. (2014) - Toll-like receptor-5 agonist Entolimod broadens the therapeutic window of 5-fluorouracil by reducing its toxicity to normal tissues in mice. In *Oncotarget* 5 (3), pp. 802–814. DOI: 10.18632/oncotarget.1773.

Köllisch, Gabriele; Kalali, Behnam Naderi; Voelcker, Verena; Wallich, Reinhard; Behrendt, Heidrun; Ring, Johannes et al. (2005) - Various members of the Toll-like receptor family contribute to the innate immune response of human epidermal keratinocytes. In *Immunology* 114 (4), pp. 531–541. DOI: 10.1111/j.1365-2567.2005.02122.x.

Kostakis, Ioannis D.; Agrogiannis, George; Vaiopoulos, Aristeidis G.; Mylona, Eleni; Patsouris, Efstratios; Kouraklis, Gregory; Koutsilieris, Michael (2015) - A clinicopathological analysis of KISS1 and KISS1R expression in colorectal cancer. In *APMIS : acta pathologica, microbiologica, et immunologica Scandinavica* 123 (7), pp. 629–637. DOI: 10.1111/apm.12397.

Kotani, M.; Detheux, M.; Vandenbogaerde, A.; Communi, D.; Vanderwinden, J. M.; Le Poul, E. et al. (2001) - The metastasis suppressor gene KiSS-1 encodes kisspeptins, the natural ligands of the orphan G protein-coupled receptor GPR54. In *The Journal of biological chemistry* 276 (37), pp. 34631–34636. DOI: 10.1074/jbc.M104847200.

Kozuki, Toshiyuki (2016) - Skin problems and EGFR-tyrosine kinase inhibitor. In *Japanese journal of clinical oncology* 46 (4), pp. 291–298. DOI: 10.1093/jjco/hyv207.

Krasinskas, Alyssa M. (2011) - EGFR Signaling in Colorectal Carcinoma. In *Pathology research international* 2011, p. 932932. DOI: 10.4061/2011/932932.

Lachmann, Peter J. (2009) - The Amplification Loop of the Complement Pathways. In Frederick W. Alt (Ed.): Advances in Immunology. Volume 104, vol. 104. 1. Aufl. s.l.: Elsevier textbooks (Advances in Immunology), pp. 115–149.

Lacouture, M. E.; Melosky, B. L. (2007) - Cutaneous reactions to anticancer agents targeting the epidermal growth factor receptor: A dermatology-oncology perspective. In *Skin therapy letter* 12 (6), pp. 1–5.

Lacouture, Mario E. (2006) - Mechanisms of cutaneous toxicities to EGFR inhibitors. In *Nature reviews. Cancer* 6 (10), pp. 803–812. DOI: 10.1038/nrc1970.

Lacouture, Mario E. (2009) - The growing importance of skin toxicity in EGFR inhibitor therapy. In *Oncology (Williston Park, N.Y.)* 23 (2), 194, 196.

Lacouture, Mario E.; Anadkat, Milan J.; Bensadoun, René-Jean; Bryce, Jane; Chan, Alexandre; Epstein, Joel B. et al. (2011) - Clinical practice guidelines for the prevention and treatment of EGFR inhibitor-associated dermatologic toxicities. In *Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer* 19 (8), pp. 1079–1095. DOI: 10.1007/s00520-011-1197-6.

Lacouture, Mario E.; Maitland, Michael L.; Segaert, Siegfried; Setser, Ann; Baran, Robert; Fox, Lindy P. et al. (2010) - A proposed EGFR inhibitor dermatologic adverse event-specific grading scale from the MASCC skin toxicity study group. In *Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer* 18 (4), pp. 509–522. DOI: 10.1007/s00520-009-0744-x.

Lamb, Justin; Crawford, Emily D.; Peck, David; Modell, Joshua W.; Blat, Irene C.; Wrobel, Matthew J. et al. (2006) - The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. In *Science (New York, N.Y.)* 313 (5795), pp. 1929–1935. DOI: 10.1126/science.1132939.

Landi, Debora; Gemignani, Federica; Pardini, Barbara; Naccarati, Alessio; Garritano, Sonia; Vodicka, Pavel et al. (2012) - Identification of candidate genes carrying polymorphisms associated with the risk of colorectal cancer by analyzing the colorectal mutome and microRNAome. In *Cancer* 118 (19), pp. 4670–4680. DOI: 10.1002/cncr.27435.

Lao, V. V.; Grady, W. M. (2012) - The Role of Timp3 in the Pathogenesis of Colorectal Cancer and Timp3 Promoter Methylation as a Potential Predictive Marker for Egfr Inhibitor Therapy. In *Journal of Surgical Research* 172 (2), p. 306. DOI: 10.1016/j.jss.2011.11.539.

Lee, J. H.; Miele, M. E.; Hicks, D. J.; Phillips, K. K.; Trent, J. M.; Weissman, B. E.; Welch, D. R. (1996) - KiSS-1, a novel human malignant melanoma metastasis-suppressor gene. In *JNCI Journal of the National Cancer Institute* 88 (23), pp. 1731–1737.

Lehninger, Albert L. (1975) - Biochemistry: The molecular basis of cell structure and function. 2. ed., 1. print. New York NY: Worth Publ.

Leporini, Christian; Saullo, Francesca; Filippelli, Gianfranco; Sorrentino, Antonio; Lucia, Maria; Perri, Gino et al. (2013) - Management of dermatologic toxicities associated with monoclonal antibody epidermal growth factor receptor inhibitors: A case review. In *Journal of pharmacology & pharmacotherapeutics* 4 (Suppl 1), S78-85. DOI: 10.4103/0976-500X.120966.

Lettmann, Sandra; Bloch, Wilhelm; Maaß, Tobias; Niehoff, Anja; Schulz, Jan-Niklas; Eckes, Beate et al. (2014) - Col6a1 null mice as a model to study skin phenotypes in patients with collagen VI related myopathies: Expression of classical and novel collagen VI variants during wound healing. In *PloS one* 9 (8), e105686. DOI: 10.1371/journal.pone.0105686.

Li, Heng; Handsaker, Bob; Wysoker, Alec; Fennell, Tim; Ruan, Jue; Homer, Nils et al. (2009a) - The Sequence Alignment/Map format and SAMtools. In *Bioinformatics (Oxford, England)* 25 (16), pp. 2078–2079. DOI: 10.1093/bioinformatics/btp352.

Li, Jiao; Zhu, Xiaoyan; Chen, Jake Yue (2009b) - Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. In *PLoS computational biology* 5 (7), e1000450. DOI: 10.1371/journal.pcbi.1000450.

Li, L.; Ying, J.; Li, H.; Zhang, Y.; Shu, X.; Fan, Y. et al. (2012) - The human cadherin 11 is a pro-apoptotic tumor suppressor modulating cell stemness through Wnt/β-catenin signaling and silenced in common carcinomas. In *Oncogene* 31 (34), pp. 3901–3912. DOI: 10.1038/onc.2011.541.

Linggi, Bryan; Carpenter, Graham (2006) - ErbB receptors: New insights on mechanisms and biology. In *Trends in cell biology* 16 (12), pp. 649–656. DOI: 10.1016/j.tcb.2006.10.008.

Lipson, Evan J.; Drake, Charles G. (2011) - Ipilimumab: an anti-CTLA-4 antibody for metastatic melanoma. In *Clinical cancer research : an official journal of the American Association for Cancer Research* 17 (22), pp. 6958–6962. DOI: 10.1158/1078-0432.CCR-11-1595.

Löffek, Stefanie; Hurskainen, Tiina; Jackow, Joanna; Sigloch, Florian Christoph; Schilling, Oliver; Tasanen, Kaisa et al. (2014) - Transmembrane collagen XVII modulates integrin dependent

keratinocyte migration via PI3K/Rac1 signaling. In *PloS one* 9 (2), e87263. DOI: 10.1371/journal.pone.0087263.

Loft, N. D.; Skov, L.; Iversen, L.; Gniadecki, R.; Dam, T. N.; Brandslund, I. et al. (2017) - Associations between functional polymorphisms and response to biological treatment in Danish patients with psoriasis. In *The pharmacogenomics journal. DOI:* 10.1038/tpj.2017.31.

LoRusso, Patricia (2009) - Toward evidence-based management of the dermatologic effects of EGFR inhibitors. In *Oncology (Williston Park, N.Y.)* 23 (2), pp. 186–194.

Lowes, Michelle A.; Suárez-Fariñas, Mayte; Krueger, James G. (2014) - Immunology of psoriasis. In *Annual review of immunology* 32, pp. 227–255. DOI: 10.1146/annurev-immunol-032713-120225.

Lupu, I.; Voiculescu, V. M.; Bacalbasa, N.; Prie, B. E.; Cojocaru, I.; Giurcaneanu, C. (2015) - Cutaneous adverse reactions specific to epidermal growth factor receptor inhibitors. In *Journal of medicine and life* 8 Spec Issue, pp. 57–61.

Lynch, Thomas J.; Bell, Daphne W.; Sordella, Raffaella; Gurubhagavatula, Sarada; Okimoto, Ross A.; Brannigan, Brian W. et al. (2004) - Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. In *The New England journal of medicine* 350 (21), pp. 2129–2139. DOI: 10.1056/NEJMoa040938.

Ma, Qiang; Lu, Anthony Y. H. (2011) - Pharmacogenetics, pharmacogenomics, and individualized medicine. In *Pharmacological reviews* 63 (2), pp. 437–459. DOI: 10.1124/pr.110.003533.

Marchler-Bauer, Aron; Bo, Yu; Han, Lianyi; He, Jane; Lanczycki, Christopher J.; Lu, Shennan et al. (2017) - CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. In *Nucleic acids research* 45 (D1), D200-D203. DOI: 10.1093/nar/gkw1129.

Margadant, Coert; Frijns, Evelyne; Wilhelmsen, Kevin; Sonnenberg, Arnoud (2008) - Regulation of hemidesmosome disassembly by growth factor receptors. In *Current opinion in cell biology* 20 (5), pp. 589–596. DOI: 10.1016/j.ceb.2008.05.001.

Marini, Joan C.; Reich, Adi; Smith, Simone M. (2014) - Osteogenesis imperfecta due to mutations in non-collagenous genes: Lessons in the biology of bone formation. In *Current opinion in pediatrics* 26 (4), pp. 500–507. DOI: 10.1097/MOP.0000000000000117.

Mariotti, A.; Kedeshian, P. A.; Dans, M.; Curatola, A. M.; Gagnoux-Palacios, L.; Giancotti, F. G. (2001) - EGF-R signaling through Fyn kinase disrupts the function of integrin alpha6beta4 at hemidesmosomes: Role in epithelial cell migration and carcinoma invasion. In *The Journal of cell biology* 155 (3), pp. 447–458. DOI: 10.1083/jcb.200105017.

Marneros, Alexander G.; Olsen, Bjorn R. (2005) - Physiological role of collagen XVIII and endostatin. In *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 19 (7), pp. 716–728. DOI: 10.1096/fj.04-2134rev.

Marsaud, Véronique; Tchakarska, Guergana; Andrieux, Geoffroy; Liu, Jian-Miao; Dembele, Doulaye; Jost, Bernard et al. (2010) - Cyclin K and cyclin D1b are oncogenic in myeloma cells. In *Molecular cancer* 9, p. 103. DOI: 10.1186/1476-4598-9-103.

Marshall, Neil A.; Galvin, Karen C.; Corcoran, Anna-Maria B.; Boon, Louis; Higgs, Rowan; Mills, Kingston H. G. (2012) - Immunotherapy with PI3K inhibitor and Toll-like receptor agonist induces IFN-γ+IL-17+ polyfunctional T cells that mediate rejection of murine tumors. In *Cancer research* 72 (3), pp. 581–591. DOI: 10.1158/0008-5472.CAN-11-0307.

Martínez-Lostao, Luis; Anel, Alberto; Pardo, Julián (2015) - How Do Cytotoxic Lymphocytes Kill Cancer Cells? In *Clinical cancer research : an official journal of the American Association for Cancer Research* 21 (22), pp. 5047–5056. DOI: 10.1158/1078-0432.CCR-15-0685.

Mastellos, D. C.; Ricklin, D.; Hajishengallis, E.; Hajishengallis, G.; Lambris, J. D. (2016) - Complement therapeutics in inflammatory diseases: promising drug candidates for C3-targeted intervention. In *Molecular oral microbiology* 31 (1), pp. 3–17. DOI: 10.1111/omi.12129.

McDougall, Steven R.; Anderson, Alexander R. A.; Chaplain, Mark A. J. (2006) - Mathematical modelling of dynamic adaptive tumour-induced angiogenesis: Clinical implications and therapeutic targeting strategies. In *Journal of theoretical biology* 241 (3), pp. 564–589. DOI: 10.1016/j.jtbi.2005.12.022.

McGowan, P. M.; Mullooly, M.; Caiazza, F.; Sukor, S.; Madden, S. F.; Maguire, A. A. et al. (2013) - ADAM-17: A novel therapeutic target for triple negative breast cancer. In *Annals of oncology : official journal of the European Society for Medical Oncology* 24 (2), pp. 362–369. DOI: 10.1093/annonc/mds279.

Mead, E. J.; Maguire, J. J.; Kuc, R. E.; Davenport, A. P. (2007) - Kisspeptins: A multifunctional peptide system with a role in reproduction, cancer and the cardiovascular system. In *British journal of pharmacology* 151 (8), pp. 1143–1153. DOI: 10.1038/sj.bjp.0707295.

Melosky, B.; Burkes, R.; Rayson, D.; Alcindor, T.; Shear, N.; Lacouture, M. (2009) - Management of skin rash during EGFR-targeted monoclonal antibody treatment for gastrointestinal malignancies: Canadian recommendations. In *Current oncology (Toronto, Ont.)* 16 (1), pp. 16–26.

Merle, Nicolas S.; Church, Sarah Elizabeth; Fremeaux-Bacchi, Veronique; Roumenina, Lubka T. (2015a) - Complement System Part I - Molecular Mechanisms of Activation and Regulation. In *Frontiers in immunology* 6, p. 262. DOI: 10.3389/fimmu.2015.00262.

Merle, Nicolas S.; Noe, Remi; Halbwachs-Mecarelli, Lise; Fremeaux-Bacchi, Veronique; Roumenina, Lubka T. (2015b) - Complement System Part II: Role in Immunity. In *Frontiers in immunology* 6, p. 257. DOI: 10.3389/fimmu.2015.00257.

Miele, Mary E.; Robertson, Gavin; Lee, Jeong-Hyung; Coleman, Aaron; McGary, Carl T.; Fisher, Paul B. et al. (1996) - Metastasis suppressed, but tumorigenicity and local invasiveness unaffected, in the human melanoma cell line MelJuSo after introduction of human chromosomes 1 or 6. In *Mol. Carcinog.* 15 (4), pp. 284–299. DOI: 10.1002/(SICI)1098-2744(199604)15:4<284::AID-MC6>3.0.CO;2-G.

Miller, Lloyd S.; Sørensen, Ole E.; Liu, Philip T.; Jalian, H. Ray; Eshtiaghpour, Deborah; Behmanesh, Behnaz E. et al. (2005) - TGF-alpha regulates TLR expression and function on epidermal keratinocytes. In *Journal of immunology (Baltimore, Md. : 1950)* 174 (10), pp. 6137–6143.

Moilanen, Jyri M.; Kokkonen, Nina; Löffek, Stefanie; Väyrynen, Juha P.; Syväniemi, Erkki; Hurskainen, Tiina et al. (2015) - Collagen XVII expression correlates with the invasion and metastasis of colorectal cancer. In *Human pathology* 46 (3), pp. 434–442. DOI: 10.1016/j.humpath.2014.11.020.

Moody, Terry W.; Nuche-Berenguer, Bernardo; Nakamura, Taichi; Jensen, Robert T. (2016) - EGFR Transactivation by Peptide G Protein-Coupled Receptors in Cancer. In *Current drug targets* 17 (5), pp. 520–528.

Moosmann, Nicolas; Weikersthal, Ludwig Fischer von; Vehling-Kaiser, Ursula; Stauch, Martina; Hass, Holger G.; Dietzfelbinger, Herrmann et al. (2011) - Cetuximab plus capecitabine and irinotecan compared with cetuximab plus capecitabine and oxaliplatin as first-line treatment for patients with metastatic colorectal cancer: AIO KRK-0104--a randomized trial of the German AIO CRC study group. In *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 29 (8), pp. 1050–1058. DOI: 10.1200/JCO.2010.31.1936.

Mordechai, Shikma; Gradstein, Libe; Pasanen, Annika; Ofir, Rivka; El Amour, Khalil; Levy, Jaime et al. (2011) - High myopia caused by a mutation in LEPREL1, encoding prolyl 3-hydroxylase 2. In *American journal of human genetics* 89 (3), pp. 438–445. DOI: 10.1016/j.ajhg.2011.08.003.

Mori, Toshiki; Anazawa, Yoshio; Matsui, Kuniko; Fukuda, Seisuke; Nakamura, Yusuke; Arakawa, Hirofumi (2002) - Cyclin K as a direct transcriptional target of the p53 tumor suppressor. In *Neoplasia (New York, N.Y.)* 4 (3), pp. 268–274. DOI: 10.1038/sj/neo/7900235.

Moriarity, Andrew; O'Sullivan, Jacintha; Kennedy, John; Mehigan, Brian; McCormick, Paul (2016) - Current targeted therapies in the treatment of advanced colorectal cancer: A review. In *Therapeutic advances in medical oncology* 8 (4), pp. 276–293. DOI: 10.1177/1758834016646734.

Moya, Patricia; Esteban, Sergio; Fernandez-Suarez, Antonio; Maestro, Marisa; Morente, Manuel; Sánchez-Carbayo, Marta (2013) - KiSS-1 methylation and protein expression patterns contribute to diagnostic and prognostic assessments in tissue specimens for colorectal cancer. In *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine* 34 (1), pp. 471–479. DOI: 10.1007/s13277-012-0572-3.

Murai, Toshiyuki (2012) - The role of lipid rafts in cancer cell adhesion and migration. In *International journal of cell biology* 2012, p. 763283. DOI: 10.1155/2012/763283.

Nakajima, Go; Patino-Garcia, Ana; Bruheim, Skjalg; Xi, Yaguang; San Julian, Mikel; Lecanda, Fernando et al. (2008) - CDH11 expression is associated with survival in patients with osteosarcoma. In *Cancer genomics & proteomics* 5 (1), pp. 37–42.

Nanney, Lillian B.; Stoscheck, Christa M.; Underwood, Robert A.; Holbrook, Karen A.; King, Lloyd E. (1990) - Immunolocalization of Epidermal Growth Factor Receptors in Normal Developing Human Skin. In *Journal of Investigative Dermatology* 94 (6), pp. 742–748. DOI: 10.1111/1523-1747.ep12874601.

Nat. Methods Editorial (2005) - Seeing is believing. In *Nat Meth* 2 (12), p. 889. DOI: 10.1038/nmeth1205-889.

Newman, M. E. J. (2006) - Modularity and community structure in networks. In *Proceedings of the National Academy of Sciences of the United States of America* 103 (23), pp. 8577–8582. DOI: 10.1073/pnas.0601602103.

NIH-NCI (2013a) - FDA Approval for Cetuximab. National Cancer Institute (US). Available online at https://www.cancer.gov/about-cancer/treatment/drugs/fda-cetuximab, updated on 7/12/2013.

NIH-NCI (2013b) - FDA Approval for Panitumumab. National Cancer Institute (US). Available online at https://www.cancer.gov/about-cancer/treatment/drugs/fda-panitumumab, updated on 7/3/2013.

NIH-NCI (2016a) - Signs and Symptoms of Colon Cancer. National Cancer Institute (US). Available online at http://www.cancer.org/cancer/news/features/signs-and-symptoms-of-colon-cancer, updated on 2/29/2016.

NIH-NCI (2016b) - Treatment of colon cancer, by stage. Edited by National Cancer Institute (US). Available online at http://www.cancer.org/cancer/colonandrectumcancer/detailedguide/colorectal-cancer-treating-by-stage-colon, updated on 1/20/2016.

NIH-NCI (2011) - FDA Approval for Gefitinib. National Cancer Institute (US). Available online at https://www.cancer.gov/about-cancer/treatment/drugs/fda-gefitinib, updated on 1/18/2011.

NIH-NCI (2017) - Common Terminology Criteria for Adverse Events (CTCAE). Available online at https://ctep.cancer.gov/protocoldevelopment/electronic_applications/ctc.htm.

Nishie, Wataru; Kiritsi, Dimitra; Nyström, Alexander; Hofmann, Silke C.; Bruckner-Tuderman, Leena (2011) - Dynamic interactions of epidermal collagen XVII with the extracellular matrix: Laminin 332 as a major binding partner. In *The American Journal of Pathology* 179 (2), pp. 829–837. DOI: 10.1016/j.ajpath.2011.04.019.

Nishie, Wataru; Natsuga, Ken; Iwata, Hiroaki; Izumi, Kentaro; Ujiie, Hideyuki; Toyonaga, Ellen et al. (2015) - Context-Dependent Regulation of Collagen XVII Ectodomain Shedding in Skin. In *The American Journal of Pathology* 185 (5), pp. 1361–1371. DOI: 10.1016/j.ajpath.2015.01.012.

Oakley, Amy E.; Clifton, Donald K.; Steiner, Robert A. (2009) - Kisspeptin signaling in the brain. In *Endocrine reviews* 30 (6), pp. 713–743. DOI: 10.1210/er.2009-0005.

Ogino, Takayuki; Nishimura, Junichi; Barman, Soumik; Kayama, Hisako; Uematsu, Satoshi; Okuzaki, Daisuke et al. (2013) - Increased Th17-inducing activity of CD14+ CD163 low myeloid cells in intestinal lamina propria of patients with Crohn's disease. In *Gastroenterology* 145 (6), 1380-91.e1. DOI: 10.1053/j.gastro.2013.08.049.

Ohhara, Yoshihito; Fukuda, Naoki; Takeuchi, Satoshi; Honma, Rio; Shimizu, Yasushi; Kinoshita, Ichiro; Dosaka-Akita, Hirotoshi (2016) - Role of targeted therapy in metastatic colorectal cancer. In *World journal of gastrointestinal oncology* 8 (9), pp. 642–655. DOI: 10.4251/wjgo.v8.i9.642.

Okugawa, Yoshinaga; Inoue, Yasuhiro; Tanaka, Koji; Toiyama, Yuji; Shimura, Tadanobu; Okigami, Masato et al. (2013) - Loss of the metastasis suppressor gene KiSS1 is associated with lymph node metastasis and poor prognosis in human colorectal cancer. In *Oncology reports* 30 (3), pp. 1449–1454. DOI: 10.3892/or.2013.2558.

Oliva, José Luis; Griner, Erin M.; Kazanietz, Marcelo G. (2005) - PKC isozymes and diacylglycerol-regulated proteins as effectors of growth factor receptors. In *Growth factors (Chur, Switzerland)* 23 (4), pp. 245–252. DOI: 10.1080/08977190500366043.

Olsen, B. R. (1997) - Collagen IX. In *The International Journal of Biochemistry & Cell Biology* 29 (4), pp. 555–558.

Ono, Keiichiro; Muetze, Tanja; Kolishovski, Georgi; Shannon, Paul; Demchak, Barry (2015) - CyREST: Turbocharging Cytoscape Access for External Tools via a RESTful API. In *F1000Research* 4, p. 478. DOI: 10.12688/f1000research.6767.1.

Osoegawa, K.; Mammoser, A. G.; Wu, C.; Frengen, E.; Zeng, C.; Catanese, J. J.; Jong, P. J. de (2001) - A bacterial artificial chromosome library for sequencing the complete human genome. In *Genome research* 11 (3), pp. 483–496. DOI: 10.1101/gr.169601.

Pabla, Baldeep; Bissonnette, Marc; Konda, Vani J. (2015) - Colon cancer and the epidermal growth factor receptor: Current treatment paradigms, the importance of diet, and the role of chemoprevention. In *World journal of clinical oncology* 6 (5), pp. 133–141. DOI: 10.5306/wjco.v6.i5.133.

Paculová, Hana; Kohoutek, Jiří (2017) - The emerging roles of CDK12 in tumorigenesis. In *Cell division* 12. DOI: 10.1186/s13008-017-0033-x.

Paez, J. Guillermo; Jänne, Pasi A.; Lee, Jeffrey C.; Tracy, Sean; Greulich, Heidi; Gabriel, Stacey et al. (2004) - EGFR mutations in lung cancer: Correlation with clinical response to gefitinib therapy. In *Science (New York, N.Y.)* 304 (5676), pp. 1497–1500. DOI: 10.1126/science.1099314.

Paila, Umadevi; Chapman, Brad A.; Kirchner, Rory; Quinlan, Aaron R. (2013) - GEMINI: Integrative exploration of genetic variation and genome annotations. In *PLoS computational biology* 9 (7), e1003153. DOI: 10.1371/journal.pcbi.1003153.

Panelius, Jaana; Meri, Seppo (2015) - Complement system in dermatological diseases - fire under the skin. In *Frontiers in medicine* 2, p. 3. DOI: 10.3389/fmed.2015.00003.

Parikka, Mataleena; Kainulainen, Tiina; Tasanen, Kaisa; Väänänen, Anu; Bruckner-Tuderman, Leena; Salo, Tuula (2003) - Alterations of collagen XVII expression during transformation of oral epithelium to dysplasia and carcinoma. In *The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society* 51 (7), pp. 921–929. DOI: 10.1177/002215540305100707.

Parsons, Jason L.; Tait, Phillip S.; Finch, David; Dianova, Irina I.; Allinson, Sarah L.; Dianov, Grigory L. (2008) - CHIP-mediated degradation and DNA damage-dependent stabilization regulate base excision repair proteins. In *Molecular cell* 29 (4), pp. 477–487. DOI: 10.1016/j.molcel.2007.12.027.

Parvizi, Javad; Kim, Gregory K. (2010) - Collagen. In Javad Parvizi (Ed.): High yield orthopaedics. 1st ed. Philadelphia: Saunders/Elsevier, pp. 107–109.

Pastrello, Chiara; Pasini, Elisa; Kotlyar, Max; Otasek, David; Wong, Serene; Sangrar, Waheed et al. (2014) - Integration, visualization and analysis of human interactome. In *Biochemical and biophysical research communications* 445 (4), pp. 757–773. DOI: 10.1016/j.bbrc.2014.01.151.

Patel, Jai N.; Wiebe, Lauren A.; Dunnenberger, Henry M.; McLeod, Howard L. (2018) - Value of Supportive Care Pharmacogenomics in Oncology Practice. In *The oncologist. DOI:* 10.1634/theoncologist.2017-0599.

Pentcheva-Hoang, Tsvetelina; Egen, Jackson G.; Wojnoonski, Kathleen; Allison, James P. (2004) - B7-1 and B7-2 selectively recruit CTLA-4 and CD28 to the immunological synapse. In *Immunity* 21 (3), pp. 401–413. DOI: 10.1016/j.immuni.2004.06.017.

Peréz-Soler, Román; Saltz, Leonard (2005) - Cutaneous adverse effects with HER1/EGFR-targeted agents: Is there a silver lining? In *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 23 (22), pp. 5235–5246. DOI: 10.1200/JCO.2005.00.6916.

Pesic, Marina; Greten, Florian R. (2016) - Inflammation and cancer: Tissue regeneration gone awry. In *Current opinion in cell biology* 43, pp. 55–61. DOI: 10.1016/j.ceb.2016.07.010.

Petrelli, F.; Borgonovo, K.; Barni, S. (2013) - The predictive role of skin rash with cetuximab and panitumumab in colorectal cancer patients: A systematic review and meta-analysis of published trials. In *Targeted oncology* 8 (3), pp. 173–181. DOI: 10.1007/s11523-013-0257-x.

Pio, Ruben; Corrales, Leticia; Lambris, John D. (2014) - The role of complement in tumor growth. In *Advances in experimental medicine and biology* 772, pp. 229–262. DOI: 10.1007/978-1-4614-5915-6_11.

Porras, Pablo; Duesbury, Margaret; Fabregat, Antonio; Ueffing, Marius; Orchard, Sandra; Gloeckner, Christian Johannes; Hermjakob, Henning (2015) - A visual review of the interactome of LRRK2: Using deep-curated molecular interaction data to represent biology. In *Proteomics* 15 (8), pp. 1390–1404. DOI: 10.1002/pmic.201400390.

Rajkumar, Thangarajan; Sabitha, Kesavan; Vijayalakshmi, Neelakantan; Shirley, Sundersingh; Bose, Mayil Vahanan; Gopal, Gopisetty; Selvaluxmy, Ganesharaja (2011) - Identification and validation of genes involved in cervical tumourigenesis. In *BMC cancer* 11, p. 80. DOI: 10.1186/1471-2407-11-80.

Rappaport, Noa; Fishilevich, Simon; Nudel, Ron; Twik, Michal; Belinky, Frida; Plaschkes, Inbar et al. (2017) - Rational confederation of genes and diseases: NGS interpretation via GeneCards, MalaCards and VarElect. In *BioMed Eng OnLine* 16 (S1), p. 119. DOI: 10.1186/s12938-017-0359-2.

Ray, Aurélie; Cot, Marlène; Puzo, Germain; Gilleron, Martine; Nigou, Jérôme (2013) - Bacterial cell wall macroamphiphiles: Pathogen-/microbe-associated molecular patterns detected by mammalian innate immune system. In *Biochimie* 95 (1), pp. 33–42. DOI: 10.1016/j.biochi.2012.06.007.

Reis, E. S.; Falcão, D. A.; Isaac, L. (2006) - Clinical aspects and molecular basis of primary deficiencies of complement component C3 and its regulatory proteins factor I and factor H. In *Scandinavian journal of immunology* 63 (3), pp. 155–168. DOI: 10.1111/j.1365-3083.2006.01729.x.

Rhee, Sang Hoon; Im, Eunok; Pothoulakis, Charalabos (2008) - Toll-like receptor 5 engagement modulates tumor development and growth in a mouse xenograft model of human colon cancer. In *Gastroenterology* 135 (2), pp. 518–528. DOI: 10.1053/j.gastro.2008.04.022.

Ricklin, Daniel; Hajishengallis, George; Yang, Kun; Lambris, John D. (2010) - Complement: a key system for immune surveillance and homeostasis. In *Nature immunology* 11 (9), pp. 785–797. DOI: 10.1038/ni.1923.

Ricklin, Daniel; Lambris, John D. (2013) - Complement in immune and inflammatory disorders: pathophysiological mechanisms. In *Journal of immunology (Baltimore, Md. : 1950)* 190 (8), pp. 3831–3838. DOI: 10.4049/jimmunol.1203487.

Ricklin, Daniel; Lambris, John D. (2016) - Therapeutic control of complement activation at the level of the central component C3. In *Immunobiology* 221 (6), pp. 740–746. DOI: 10.1016/j.imbio.2015.06.012.

Robert, Caroline; Soria, Jean-Charles; Spatz, Alain; Le Cesne, Axel; Malka, David; Pautier, Patricia et al. (2005) - Cutaneous side-effects of kinase inhibitors and blocking antibodies. In *The Lancet Oncology* 6 (7), pp. 491–500. DOI: 10.1016/S1470-2045(05)70243-6.

Robert Koch-Institut (2015) - Krebs in Deutschland 2011/2012. Gesundheitsberichterstattung des Bundes. 10 Ausg. 2015.

Robinson, D. R.; Wu, Y. M.; Lin, S. F. (2000) - The protein tyrosine kinase family of the human genome. In *Oncogene* 19 (49), pp. 5548–5557. DOI: 10.1038/sj.onc.1203957.

Row, Sindhu; Liu, Yayu; Alimperti, Stella; Agarwal, Sandeep K.; Andreadis, Stelios T. (2016) - Cadherin-11 is a novel regulator of extracellular matrix synthesis and tissue mechanics. In *Journal of cell science* 129 (15), pp. 2950–2961. DOI: 10.1242/jcs.183772.

Saladzinskas, Zilvinas; Tamelis, Algimantas; Paskauskas, Saulius; Pranys, Darius; Pavalkis, Dainius (2010) - Facial skin metastasis of colorectal cancer: A case report. In *Cases journal* 3, p. 28. DOI: 10.1186/1757-1626-3-28.

Scardoni, Giovanni; Tosadori, Gabriele; Pratap, Sakshi; Spoto, Fausto; Laudanna, Carlo (2015) - Finding the shortest path with PesCa: a tool for network reconstruction. In *F1000Research* 4, p. 484. DOI: 10.12688/f1000research.6769.2.

Schaaf, Sebastian; Nazeer Batcha, Aarif Mohamed; Zhang, Guokun; Fischer, Sandra; Varadharajan, Ashok; Mansmann, Robert (2014) - The Munich NGS-FabLab: A glimpse on an IT infrastructure for medical sequence data. Baltimore, USA (Poster Session Galaxy Community Conference (GCC) 2014,

P10). Available online at https://depot.galaxyproject.org/hub/attachments/documents/posters/gcc2014/P10Schaaf.pdf.

Schecher, Sabrina; Walter, Britta; Falkenstein, Michael; Macher-Goeppinger, Stephan; Stenzel, Philipp; Krümpelmann, Kristina et al. (2017) - Cyclin K dependent regulation of Aurora B affects apoptosis and proliferation by induction of mitotic catastrophe in prostate cancer. In *International journal of cancer* 141 (8), pp. 1643–1653. DOI: 10.1002/ijc.30864.

Schlessinger, Joseph (2002) - Ligand-Induced, Receptor-Mediated Dimerization and Activation of EGF Receptor. In *Cell* 110 (6), pp. 669–672. DOI: 10.1016/S0092-8674(02)00966-2.

Schmiegel, W.; Pox, C.; Reinacher-Schick, A.; Adler, G.; Fleig, W.; Fölsch, U. R. et al. (2008) - S3-Leitlinie "Kolorektales Karzinom". In *Z Gastroenterol* (46), pp. 1–73. DOI: 10.1055/s-2008-1027700.

Sengupta, Pritam K.; Smith, Erin M.; Kim, Kwonseop; Murnane, Mary Jo; Smith, Barbara D. (2003) - DNA hypermethylation near the transcription start site of collagen alpha2(I) gene occurs in both cancer cell lines and primary colorectal cancers. In *Cancer research* 63 (8), pp. 1789–1797.

Seppinen, Lotta; Pihlajaniemi, Taina (2011) - The multiple functions of collagen XVIII in development and disease. In *Matrix biology : journal of the International Society for Matrix Biology* 30 (2), pp. 83–92. DOI: 10.1016/j.matbio.2010.11.001.

Seshacharyulu, Parthasarathy; Ponnusamy, Moorthy P.; Haridas, Dhanya; Jain, Maneesh; Ganti, Apar K.; Batra, Surinder K. (2012) - Targeting the EGFR signaling pathway in cancer therapy. In *Expert opinion on therapeutic targets* 16 (1), pp. 15–31. DOI: 10.1517/14728222.2011.648617.

Shackelford, David B.; Shaw, Reuben J. (2009) - The LKB1-AMPK pathway: Metabolism and growth control in tumour suppression. In *Nature reviews. Cancer* 9 (8), pp. 563–575. DOI: 10.1038/nrc2676.

Shah, R.; Smith, P.; Purdie, C.; Quinlan, P.; Baker, L.; Aman, P. et al. (2009) - The prolyl 3-hydroxylases P3H2 and P3H3 are novel targets for epigenetic silencing in breast cancer. In *British journal of cancer* 100 (10), pp. 1687–1696. DOI: 10.1038/sj.bjc.6605042.

Shang, Yu; Xu, Xialian; Duan, Xiaolin; Guo, Junwei; Wang, Yinyin; Ren, Fangli et al. (2014) - Hsp70 and Hsp90 oppositely regulate TGF-β signaling through CHIP/Stub1. In *Biochemical and biophysical research communications* 446 (1), pp. 387–392. DOI: 10.1016/j.bbrc.2014.02.124.

Shannon, Paul; Markiel, Andrew; Ozier, Owen; Baliga, Nitin S.; Wang, Jonathan T.; Ramage, Daniel et al. (2003) - Cytoscape: A software environment for integrated models of biomolecular interaction networks. In *Genome research* 13 (11), pp. 2498–2504. DOI: 10.1101/gr.1239303.

Shaykhiev, Renat; Behr, Jürgen; Bals, Robert (2008) - Microbial patterns signaling via Toll-like receptors 2 and 5 contribute to epithelial repair, growth and survival. In *PloS one* 3 (1), e1393. DOI: 10.1371/journal.pone.0001393.

Sherwani, Mohammad Asif; Tufail, Saba; Muzaffar, Anum Fatima; Yusuf, Nabiha (2017) - The skin microbiome and immune system: Potential target for chemoprevention? In *Photodermatology, photoimmunology & photomedicine. DOI:* 10.1111/phpp.12334.

Shi, Qiaoni; Chen, Ye-Guang (2017) - Interplay between TGF-β signaling and receptor tyrosine kinases in tumor development. In *Science China. Life sciences* 60 (10), pp. 1133–1141. DOI: 10.1007/s11427-017-9173-5.

Siegel, Rebecca; Naishadham, Deepa; Jemal, Ahmedin (2012) - Cancer statistics, 2012. In *CA: a cancer journal for clinicians* 62 (1), pp. 10–29. DOI: 10.3322/caac.20138.

Slater, Ted (2014) - Recent advances in modeling languages for pathway maps and computable biological networks. In *Drug discovery today* 19 (2), pp. 193–198. DOI: 10.1016/j.drudis.2013.12.011.

Smith, Heath A.; Kang, Yibin (2013) - The metastasis-promoting roles of tumor-associated immune cells. In *Journal of molecular medicine (Berlin, Germany)* 91 (4), pp. 411–429. DOI: 10.1007/s00109-013-1021-5.

Soo, Ross A.; Kim, Hye Ryun; Asuncion, Bernadette Reyna; Fazreen, Zul; Omar, Mohamed Feroz Mohd; Herrera, Maria Cynthia et al. (2017) - Significance of immune checkpoint proteins in EGFR-mutant non-small cell lung cancer. In *Lung cancer (Amsterdam, Netherlands)* 105, pp. 17–22. DOI: 10.1016/j.lungcan.2017.01.008.

Steinmann, Beat; Royce, Peter M.; Superti-Furga, Andrea (2003) - The Ehlers-Danlos Syndrome. In Peter M. Royce, Beat Steinmann (Eds.): Connective tissue and its heritable disorders. Molecular, genetic, and medical aspects. 2nd ed. New York, Chichester: Wiley, pp. 431–523.

Stuart, Darrin; Sellers, William R. (2009) - Linking somatic genetic alterations in cancer to therapeutics. In *Current opinion in cell biology* 21 (2), pp. 304–310. DOI: 10.1016/j.ceb.2009.02.001.

Sukumar, N.; Krein, Michael P. (2012) - Graphs and networks in chemical and biological informatics: Past, present and future. In *Future medicinal chemistry* 4 (16), pp. 2039–2047. DOI: 10.4155/fmc.12.128.

Takino, Takahisa; Koshikawa, Naohiko; Miyamori, Hisashi; Tanaka, Motohiro; Sasaki, Takuma; Okada, Yasunori et al. (2003) - Cleavage of metastasis suppressor gene product KiSS-1 protein/metastin by matrix metalloproteinases. In *Oncogene* 22 (30), pp. 4617–4626. DOI: 10.1038/sj.onc.1206542.

Tan, Adrian; Abecasis, Gonçalo R.; Kang, Hyun Min (2015) - Unified representation of genetic variants. In *Bioinformatics (Oxford, England)* 31 (13), pp. 2202–2204. DOI: 10.1093/bioinformatics/btv112.

Tanida, Satoshi; Joh, Takashi; Itoh, Keisuke; Kataoka, Hiromi; Sasaki, Makoto; Ohara, Hirotaka et al. (2004) - The mechanism of cleavage of EGFR ligands induced by inflammatory cytokines in gastric cancer cells. In *Gastroenterology* 127 (2), pp. 559–569. DOI: 10.1053/j.gastro.2004.05.017.

Tanjore, Harikrishna; Kalluri, Raghu (2006) - The role of type IV collagen and basement membranes in cancer progression and metastasis. In *The American Journal of Pathology* 168 (3), pp. 715–717. DOI: 10.2353/ajpath.2006.051321.

Teles, Milena Gurgel; Silveira, Leticia Ferreira Gontijo; Tusset, Cintia; Latronico, Ana Claudia (2011) - New genetic factors implicated in human GnRH-dependent precocious puberty: The role of kisspeptin system. In *Molecular and cellular endocrinology* 346 (1-2), pp. 84–90. DOI: 10.1016/j.mce.2011.05.019.

Tjandra, Joe J.; Chan, Miranda K. Y. (2007) - Follow-up after curative resection of colorectal cancer: A meta-analysis. In *Diseases of the colon and rectum* 50 (11), pp. 1783–1799. DOI: 10.1007/s10350-007-9030-5.

Tng, Eng Loon (2015) - Kisspeptin signalling and its roles in humans. In *Singapore medical journal* 56 (12), pp. 649–656. DOI: 10.11622/smedj.2015183.

Torres, Sofia; Bartolomé, Rubén A.; Mendes, Marta; Barderas, Rodrigo; Fernandez-Aceñero, M. Jesús; Peláez-García, Alberto et al. (2013) - Proteome profiling of cancer-associated fibroblasts identifies novel proinflammatory signatures and prognostic markers for colorectal cancer. In *Clinical cancer research : an official journal of the American Association for Cancer Research* 19 (21), pp. 6006–6019. DOI: 10.1158/1078-0432.CCR-13-1130.

Trotti, Andy; Colevas, A. Dimitrios; Setser, Ann; Rusch, Valerie; Jaques, David; Budach, Volker et al. (2003) - CTCAE v3.0: Development of a comprehensive grading system for the adverse effects of cancer treatment. In *Seminars in radiation oncology* 13 (3), pp. 176–181. DOI: 10.1016/S1053-4296(03)00031-6.

Truedsson, Lennart; Bengtsson, Anders A.; Sturfelt, Gunnar (2007) - Complement deficiencies and systemic lupus erythematosus. In *Autoimmunity* 40 (8), pp. 560–566. DOI: 10.1080/08916930701510673.

Tsukamoto, Hiroshi; Horiuchi, Takahiko; Kokuba, Hisashi; Nagae, Shonosuke; Nishizaka, Hiroaki; Sawabe, Takuya et al. (2005) - Molecular analysis of a novel hereditary C3 deficiency with systemic

lupus erythematosus. In *Biochemical and biophysical research communications* 330 (1), pp. 298–304. DOI: 10.1016/j.bbrc.2005.02.159.

Ujiie, Hideyuki; Shibaki, Akihiko; Nishie, Wataru; Shimizu, Hiroshi (2010) - What's new in bullous pemphigoid. In *The Journal of dermatology* 37 (3), pp. 194–204. DOI: 10.1111/j.1346-8138.2009.00792.x.

van den Bergh, Françoise; Eliason, Steven L.; Burmeister, Brian T.; Giudice, George J. (2012) - Collagen XVII (BP180) modulates keratinocyte expression of the proinflammatory chemokine, IL-8. In *Experimental dermatology* 21 (8), pp. 605–611. DOI: 10.1111/j.1600-0625.2012.01529.x.

Veitia, Reiner A. (2004) - Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. In *Genetics* 168 (1), pp. 569–574. DOI: 10.1534/genetics.104.029785.

Vihinen, Pia; Paija, Outi; Kivisaari, Atte; Koulu, Leena; Aho, Heikki (2011) - Cutaneous lupus erythematosus after treatment with paclitaxel and bevacizumab for metastatic breast cancer: a case report. In *Journal of medical case reports* 5, p. 243. DOI: 10.1186/1752-1947-5-243.

Villaveces, Jose M.; Koti, Prasanna; Habermann, Bianca H. (2015) - Tools for visualization and analysis of molecular networks, pathways, and -omics data. In *Advances and applications in bioinformatics and chemistry : AABC* 8, pp. 11–22. DOI: 10.2147/AABC.S63534.

Walko, Gernot; Castañón, Maria J.; Wiche, Gerhard (2015) - Molecular architecture and function of the hemidesmosome. In *Cell and tissue research* 360 (3), pp. 529–544. DOI: 10.1007/s00441-015-2216-6.

Wang, Shuai; Li, Yi; Hu, Yun-Hong; Song, Ren; Gao, Yan; Liu, Hai-Yun et al. (2013) - STUB1 is essential for T-cell activation by ubiquitinating CARMA1. In *European journal of immunology* 43 (4), pp. 1034–1041. DOI: 10.1002/eji.201242554.

Wang, Tianxiao; Yang, Jingxuan; Xu, Jianwei; Li, Jian; Cao, Zhe; Zhou, Li et al. (2014a) - CHIP is a novel tumor suppressor in pancreatic cancer through targeting EGFR. In *Oncotarget* 5 (7), pp. 1969–1986. DOI: 10.18632/oncotarget.1890.

Wang, Yangmeng; Ren, Fangli; Wang, Yinyin; Feng, Yarui; Wang, Dianjun; Jia, Baoqing et al. (2014b) - CHIP/Stub1 functions as a tumor suppressor and represses NF-κB-mediated signaling in colorectal cancer. In *Carcinogenesis* 35 (5), pp. 983–991. DOI: 10.1093/carcin/bgt393.

Wang, Zhixiang (2016) - Transactivation of Epidermal Growth Factor Receptor by G Protein-Coupled Receptors: Recent Progress, Challenges and Future Research. In *International journal of molecular sciences* 17 (1). DOI: 10.3390/ijms17010095.

Webb, Eika S.; Liu, Peng; Baleeiro, Renato; Lemoine, Nicholas R.; Yuan, Ming; Wang, Yao-He (2017) - Immune checkpoint inhibitors in cancer therapy. In *Journal of biomedical research. DOI:* 10.7555/JBR.31.20160168.

Weis, Mary Ann; Hudson, David M.; Kim, Lammy; Scott, Melissa; Wu, Jiann-Jiu; Eyre, David R. (2010) - Location of 3-hydroxyproline residues in collagen types I, II, III, and V/XI implies a role in fibril supramolecular assembly. In *The Journal of biological chemistry* 285 (4), pp. 2580–2590. DOI: 10.1074/jbc.M109.068726.

Wells, Alan (1999) - EGF receptor. In *The International Journal of Biochemistry & Cell Biology* 31 (6), pp. 637–643. DOI: 10.1016/S1357-2725(99)00015-1.

Wells, Alan (Ed.) (2006) - Cell Motility in Cancer Invasion and Metastasis. Dordrecht: Springer (Cancer Metastasis - Biology and Treatment, 8). Available online at http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10144324.

Wen, Feng; Li, Qiu (2016) - Treatment dilemmas of cetuximab combined with chemotherapy for metastatic colorectal cancer. In *World journal of gastroenterology* 22 (23), pp. 5332–5341. DOI: 10.3748/wjg.v22.i23.5332.

West, A.; Vojta, P. J.; Welch, D. R.; Weissman, B. E. (1998) - Chromosome localization and genomic structure of the KiSS-1 metastasis suppressor gene (KISS1). In *Genomics* 54 (1), pp. 145–148. DOI: 10.1006/geno.1998.5566.

West, Alison; Jenkins, Brendan (2015) - Inflammatory and Non-Inflammatory Roles for Toll-Like Receptors in Gastrointestinal Cancer. In *CPD* 21 (21), pp. 2968–2977. DOI: 10.2174/1381612821666150514104411.

White, J. A.; McAlpine, P. J.; Antonarakis, S.; Cann, H.; Eppig, J. T.; Frazer, K. et al. (1997) - Guidelines for human gene nomenclature (1997). HUGO Nomenclature Committee. In *Genomics* 45 (2), pp. 468–471.

Willett, Christopher G.; Boucher, Yves; Di Tomaso, Emmanuelle; Duda, Dan G.; Munn, Lance L.; Tong, Ricky T. et al. (2004) - Direct evidence that the VEGF-specific antibody bevacizumab has antivascular effects in human rectal cancer. In *Nature medicine* 10 (2), pp. 145–147. DOI: 10.1038/nm988.

Wu, Guanming; Dawson, Eric; Duong, Adrian; Haw, Robin; Stein, Lincoln (2014) - ReactomeFIViz: A Cytoscape app for pathway and network-based data analysis. In *F1000Research* 3, p. 146. DOI: 10.12688/f1000research.4431.2.

Wu, Guanming; Feng, Xin; Stein, Lincoln (2010) - A human functional protein interaction network and its application to cancer data analysis. In *Genome biology* 11 (5), R53. DOI: 10.1186/gb-2010-11-5-r53.

Xu, Cheng-Wei; Zhang, Tian-Peng; Wang, Hong-Xia; Yang, Hui; Li, Hui-Hua (2013) - CHIP enhances angiogenesis and restores cardiac function after infarction in transgenic mice. In *Cellular physiology and biochemistry : international journal of experimental cellular physiology, biochemistry, and pharmacology* 31 (2-3), pp. 199–208. DOI: 10.1159/000343361.

Xu, Guangru; Zhang, Minghui; Zhu, Hongxing; Xu, Jinhua (2017) - A 15-gene signature for prediction of colon cancer recurrence and prognosis based on SVM. In *Gene* 604, pp. 33–40. DOI: 10.1016/j.gene.2016.12.016.

Yan, C.; Wang, H.; Boyd, D. D. (2001) - KiSS-1 represses 92-kDa type IV collagenase expression by down-regulating NF-kappa B binding to the promoter as a consequence of Ikappa Balpha -induced block of p65/p50 nuclear translocation. In *The Journal of biological chemistry* 276 (2), pp. 1164–1172. DOI: 10.1074/jbc.M008681200.

Yarom, Nirit; Jonker, Derek J. (2011) - The role of the epidermal growth factor receptor in the mechanism and treatment of colorectal cancer. In *Discovery medicine* 11 (57), pp. 95–105.

Yazdi, Mohammad Hossein; Faramarzi, Mohammad Ali; Nikfar, Shekoufeh; Abdollahi, Mohammad (2015) - A Comprehensive Review of Clinical Trials on EGFR Inhibitors Such as Cetuximab and Panitumumab as Monotherapy and in Combination for Treatment of Metastatic Colorectal Cancer. In *Avicenna journal of medical biotechnology* 7 (4), pp. 134–144.

Yin, Jia; Yu, Fu-Shin X. (2009) - ERK1/2 mediate wounding- and G-protein-coupled receptor ligands-induced EGFR activation via regulating ADAM17 and HB-EGF shedding. In *Investigative ophthalmology & visual science* 50 (1), pp. 132–139. DOI: 10.1167/iovs.08-2246.

Younesi, Erfan; Toldo, Luca; Müller, Bernd; Friedrich, Christoph M.; Novac, Natalia; Scheer, Alexander et al. (2012) - Mining biomarker information in biomedical literature. In *BMC medical informatics and decision making* 12, p. 148. DOI: 10.1186/1472-6947-12-148.

Yu, Hongmei; Zhou, Xiangdong; Wen, Sha; Xiao, Qian (2012) - Flagellin/TLR5 responses induce mucus hypersecretion by activating EGFR via an epithelial cell signaling cascades. In *Experimental cell research* 318 (6), pp. 723–731. DOI: 10.1016/j.yexcr.2011.12.016.

Zajac, Mateusz; Law, Jeffrey; Cvetkovic, Dragana Donna; Pampillo, Macarena; McColl, Lindsay; Pape, Cynthia et al. (2011) - GPR54 (KISS1R) transactivates EGFR to promote breast cancer cell invasiveness. In *PloS one* 6 (6), e21599. DOI: 10.1371/journal.pone.0021599.

Zhang, Z.; DuBois, R. N. (2001) - Detection of differentially expressed genes in human colon carcinoma cells treated with a selective COX-2 inhibitor. In *Oncogene* 20 (33), pp. 4450–4456. DOI: 10.1038/sj.onc.1204588.

Zhu, Qiangqiang; Wang, Zhen; Zhou, Lihua; Ren, Yan; Gong, Ying; Qin, Wei et al. (2018) - The role of cadherin-11 in microcystin-LR-induced migration and invasion in colorectal carcinoma cells. In *Oncology letters* 15 (2), pp. 1417–1422. DOI: 10.3892/ol.2017.7458.

Zimina, Elena P.; Bruckner-Tuderman, Leena; Franzke, Claus-Werner (2005) - Shedding of collagen XVII ectodomain depends on plasma membrane microenvironment. In *The Journal of biological chemistry* 280 (40), pp. 34019–34024. DOI: 10.1074/jbc.M503751200.

Zimina, Elena P.; Fritsch, Anja; Schermer, Bernhard; Bakulina, Anastasia Yu; Bashkurov, Mikhail; Benzing, Thomas; Bruckner-Tuderman, Leena (2007) - Extracellular phosphorylation of collagen XVII by ecto-casein kinase 2 inhibits ectodomain shedding. In *The Journal of biological chemistry* 282 (31), pp. 22737–22746. DOI: 10.1074/jbc.M701937200.

Zipfel, Peter F. (2001) - Complement: Alternative Pathway. In : Encyclopedia of life sciences 2001. London, New York, Vols. 21-32, Chichester, West Sussex, U.K.: Nature Pub. Group; Wiley.

7. Appendix

Tab. A1 - BLAST results for finding prolyl 3-hydroxylation motifs from collagens type I and II in collagen XVII's sequence. With GLPGPIGPPGPR and GIPGPIGPPGPR being queries and the full amino acid sequence of collagen XVII being the subject (database), in the latter eleven unique, although partly overlapping hits could detected with an expect (e-value) <= 0.01. Consequently, the proline residues highlighted in red are supposed to by potential 3Hyp sites and target to a referring P3H, presumably P3H3. Context to the overall sequence depicted in Box A1.

Match				score	expect	identities
Query	3	PGPIGPPGPR	12	27.8 bits(58)	2e-04	9/10
		PGP GPPGPR				
Sbjct	1217	PGPPGPPGPR	1226			
Query	2	IPGPIGPPGP	11	27.8 bits(58)	2e-04	9/10
		IPGP GPPGP				
Sbjct	1213	IPGPPGPPGP	1222			
Query	1	GLPGPIGPPG	10	26.5 bits(55)	6e-04	9/10
		GLPGP GPPG				
Sbjct	885	GLPGPPGPPG	894			
Sbjct	1146	GLPGPPGPPG	1155			
Query	2	LPGPIGPPGPR	12	26.1 bits(54)	9e-04	9/11
		L GP GPPGPR				
Sbjct	855	LQGPPGPPGPR	865			
Query	3	PGPIGPPGP	11	25.7 bits(53)	0.001	8/9
		PGP GPPGP				
Sbjct	673	PGPQGPPGP	681			
Sbjct	857	GPPGPPGPR	865			
Sbjct	862	PGPRGPPGP	870			
Sbjct	908	PGPPGPPGP	916			
Sbjct	995	PGPPGPPGP	1003			
Sbjct	1038	PGPPGPPGP	1046			
Query	3	PGPIGPPGP	11	24.4 bits(50)	0.004	8/9
		PGP GPPGP				
Sbjct	908	PGPPGPPGP	916			
Sbjct	995	PGPPGPPGP	1003			
Sbjct	1038	PGPPGPPGP	1046			

Box A1 - Protein sequence of collagen XVII. Grey = transmembrane domain, violet = recognition site for ADAM17, red = recognition site for CK2 (with bold 'S' marking serine dedicated to phosphorylation), green = collagenous domains. Bold underlined subsequences in collagenous domains mark putative 3'-hydroxylation patterns by similarity via BLAST (comp. Tab. A1), with double underlining or additional character strike out indicating for two or three overlapping patterns, respectively. Red 'P's mark the proline residues dedicated to 3'-hydroxylation (3Hyp positions). Sequence and domains according to Giudice *et al.* 1992, recognition sites according to Zimina *et al.* 2007.

```
1     MDVTKKNKRDGTEVTERIVTETVTTRLTSLPPKGGTSNGYAKTASLGGGSRLEKQSLTHG   60
61    SSGYINSTGSTRGHASTSSYRRAHSPASTLPNSPGSTFERKTHVTRHAYEGSSSGNSSPE  120
121   YPRKEFASSSTRGRSQTRESEIRVRLQSASPSTRWTELDDVKRLLKGSRSASVSPTRNSS  180
181   NTLPIPKKGTVETKIVTASSQSVSGTYDATILDANLPSHVWSSTLPAGSSMGTYHNNMTT  240
241   QSSSLLNTNAYSAGSVFGVPNNMASCSPTLHPGLSTSSSVFGMQNNLAPSLTTLSHGTTT  300
301   TSTAYGVKKNMPQSPAAVNTGVSTSAACTTSVQSDDLLHKDCKFLILEKDNTPAKKEMEL  360
361   LIMTKDSGKVFTASPASIAATSFSEDTLKKEKQAAYNADSGLKAEANGDLKTVSTKGKTT  420
421   TADIHSYGSSGGGGSGGGGGVGGAGGGPWGPAPAWCPCGSCCSWWKWLLGLLLTWLLLLG  480
481   LLFGLIALAEEVRKLKARVDELERIRRSILPYGDSMDRIEKDRLQGMAPAAGADLDKIGL  540
541   HSDSQEELWMFVRKKLMMEQENGNLRGSPGPKGDMGSPGPKGDRGFPGTPGIPGPLGHPG  600
601   PQGPKGQKGSVGDPGMEGPMGQRGREGPMGPRGEAGPPGSGEKGERGAAGEPGPHGPPGV  660
661   PGSVGPKGSSGSPGPQGPPGPVGLQGLRGEVGLPGVKGDKGPMGPPGPKGDQGEKGPRGL  720
721   TGEPGMRGLPGAVGEPGAKGAMGPAGPDGHQGPRGEQGLTGMPGIRGPPGPSGDPGKPGL  780
781   TGPQGPQGLPGTPGRPGIKGEPGAPGKIVTSEGSSMLTVPGPPGPPGAMGPPGPPGAPGP  840
841   AGPAGLPGHQEVLNLQGPPGPPGPRGPPGPSIPGPPGPRGPPGEGLPGPPGPPGSFLSNS  900
901   ETFLSGPPGPPGPPGPKGDQGPPGPRGHQGEQGLPGFSTSGSSSFGLNLQGPPGPPGPQG  960
961   PKGDKGDPGVPGALGIPSGPSEGGSSSTMYVSGPPGPPGPPGPSISSSGQEIQQYISE  1020
1021  YMQSDSIRSYLSGVQGPPGPPGPPGPVTTITGETFDYSELASHVVSYLRTSGYGVSLFSS  1080
1081  SISSEDILAVLQRDDVRQYLRQYLMGPRGPPGPPGASGDGSLLSLDYAELSSRILSYMSS  1140
1141  SGISIGLPGPPGPPGLPGTSYEELLSLLRGSEFRGIVGPPGPPGPPGIPGNVWSSISVED  1200
1201  LSSYLHTAGLSFIPGPPGPPGPPGPPGPRGPPGVSGALATYAAENSDSFRSELISYLTSPDVR  1260
1261  SFIVGPPGPPGPQGPPGDSRLLSTDASHSRGSSSSSHSSSVRRGSSYSSSMSTGGGGAGS  1320
1321  LGAGGAFGEAAGDRGPYGTDIGPGGGYGAAAEGGMYAGNGGLLGADFAGDLDYNELAVRV  1380
1381  SESMQRQGLLQGMAYTVQGPPGQPGPQGPPGISKVFSAYSNVTADLMDFFQTYGAIQGPP  1440
1441  GQKGEMGTPGPKGDRGPAGPPGHPGPPGPRGHKGEKGDKGDQVYAGRRRRRSIAVKP     1497
```

Box A2 - Short read data processing steps as executed by Galaxy (1). Pathes indicated by '*[xyz_path]*' are system-specific. Dataset names are arbitrary, but refer to the workflow "036_new" in the 'NGS-FabLab' Galaxy instance [Schaaf *et al.* 2014].

FASTQ Quality Trimmer: 1x concat'ed forward + 1x concat'ed reverse reads

```
python [repo_tool_path]/fastq_trimmer_by_quality.py
    '[dataset_path]/dataset_108445.dat'
    '[dataset_path]/dataset_147595.dat' -f 'sanger' -s '10' -t '1' -e
    '53' -a 'mean' -x '0' -c '>=' -q '20.0' -k
```

bwa-mem v0.7.10: paired-end (forward + reverse reads)

```
bwa mem -t "${GALAXY_SLOTS:-7}" -v 1 -T "30" -h "5" -M -R
    '@RG\tID:120410_HWUSI-
    EAS632R_00054.3\tSM:036\tPL:ILLUMINA\tLB:Fire2_CIOX\tCN:LAFUGA\tDS:S
    R negative, normal sample\tDT:2012-04-10\tPU:HWUSI-EAS632R'
    "[genome_path]/hg19.fa" "[dataset_path]/dataset_109015.dat"
    "[dataset_path]/dataset_109016.dat" | samtools view -Sb - >
    temporary_bam_file.bam && samtools sort -f temporary_bam_file.bam
    [dataset_path]/dataset_109017.dat
```

Picard v1.136: ReorderSam

```
_JAVA_OPTIONS=${_JAVA_OPTIONS:-'-Xmx2048m -Xms256m -
    XX:ParallelGCThreads=7'} && export _JAVA_OPTIONS && ln -s "hg19"
    "localref.fa" && java -jar $JAVA_JAR_PATH/picard.jar ReorderSam
    INPUT="[dataset_path]/dataset_147597.dat"
    OUTPUT="[dataset_path]/dataset_147598.dat"
    REFERENCE="[genome_path]/hg19.fa"
    ALLOW_INCOMPLETE_DICT_CONCORDANCE="false"
    ALLOW_CONTIG_LENGTH_DISCORDANCE="false"
    VALIDATION_STRINGENCY="SILENT" QUIET=true VERBOSITY=ERROR
```

Picard v1.136: MarkDuplicates

```
_JAVA_OPTIONS=${_JAVA_OPTIONS:-'-Xmx2048m -Xms256m -
    XX:ParallelGCThreads=7'} && export _JAVA_OPTIONS && java -jar
    $JAVA_JAR_PATH/picard.jar MarkDuplicates
    INPUT="[dataset_path]/dataset_147598.dat"
    OUTPUT="[dataset_path]/dataset_147600.dat"
    METRICS_FILE="[dataset_path]/dataset_147599.dat"
    REMOVE_DUPLICATES="false" ASSUME_SORTED="true"
    DUPLICATE_SCORING_STRATEGY="SUM_OF_BASE_QUALITIES"
    READ_NAME_REGEX='[a-zA-Z0-9]+:[0-9]:([0-9]+):([0-9]+):([0-9]+).*.'
    OPTICAL_DUPLICATE_PIXEL_DISTANCE="100"
    VALIDATION_STRINGENCY="SILENT" QUIET=true VERBOSITY=ERROR
```

GATK2 v2.8.0: Realigner Target Creator

```
python [repo_tool_path]/gatk2/gatk2_wrapper.py --stdout
    "[dataset_path]/dataset_147602.dat" -d "-I"
    "[dataset_path]/dataset_147600.dat" "bam" "gatk_input" -d ""
    "[metadata_path]/metadata_16756.dat" "bam_index" "gatk_input" -p '
    java -jar "$GATK2_PATH/GenomeAnalysisTK.jar" -T
    "RealignerTargetCreator" -o "[dataset_path]/dataset_147601.dat"
    $GATK2_SITE_OPTIONS --num_cpu_threads_per_data_thread 1 --
    num_threads ${GALAXY_SLOTS:-7} -R "[genome_path]/hg19.fa" '
```

Box A3 - Short read data processing steps as executed by Galaxy (2). Continued from Box A2.

GATK2 v2.8.0: Indel Realigner

```
python [repo_tool_path]/gatk2/gatk2_wrapper.py --stdout
    "[dataset_path]/dataset_147604.dat" -d "-I"
    "[dataset_path]/dataset_147600.dat" "bam" "gatk_input" -d ""
    "[metadata_path]/metadata_16756.dat" "bam_index" "gatk_input" -p '
    java -jar "$GATK2_PATH/GenomeAnalysisTK.jar" -T "IndelRealigner" -o
    "[dataset_path]/dataset_147603.dat" $GATK2_SITE_OPTIONS --
    num_cpu_threads_per_data_thread 1 -R "[genome_path]/hg19.fa" -LOD
    "5.0" ' -p '--pedigreeValidationType "STRICT"' -p ' --
    filter_bases_not_stored --filter_mismatching_base_and_quals ' -p '--
    interval_set_rule "UNION"' -p '--interval_padding "0"' -p '--
    downsampling_type "NONE"' -p ' --baq "OFF" --baqGapOpenPenalty
    "40.0" --defaultBaseQualities "-1" --validation_strictness "SILENT"
    --interval_merging "ALL" ' -d "-targetIntervals"
    "[dataset_path]/dataset_147601.dat" "gatk_interval"
    "gatk_target_intervals" -p ' --disable_bam_indexing '
```

GATK2 v2.8.0: Base Recalibrator

```
python [repo_tool_path]/gatk2/gatk2_wrapper.py --stdout
    "[dataset_path]/dataset_147606.dat" -d "-I"
    "[dataset_path]/dataset_147603.dat" "bam" "gatk_input" -d ""
    "[metadata_path]/metadata_16762.dat" "bam_index" "gatk_input" -p '
    java -jar "$GATK2_PATH/GenomeAnalysisTK.jar" -T "BaseRecalibrator"
    $GATK2_SITE_OPTIONS --num_cpu_threads_per_data_thread
    ${GALAXY_SLOTS:-7} --no_standard_covs -R "[genome_path]/hg19.fa" --
    out "[dataset_path]/dataset_147605.dat" -cov "ContextCovariate" -cov
    "CycleCovariate" ' -p '--
    run_without_dbsnp_potentially_ruining_quality'
```

GATK2 v2.8.0: Haplotype Caller

```
python [repo_tool_path]/gatk2/gatk2_wrapper.py --stdout
    "[dataset_path]/dataset_147608.dat" -d "-I"
    "[dataset_path]/dataset_147603.dat" "bam" "gatk_input_0" -d ""
    [metadata_path]/metadata_16762.dat" "bam_index" "gatk_input_0" -p '
    java -jar "$GATK2_PATH/GenomeAnalysisTK.jar" -T "HaplotypeCaller" -o
    "[dataset_path]/dataset_147607.dat" $GATK2_SITE_OPTIONS --
    num_threads ${GALAXY_SLOTS:-7} -R "[genome_path]/hg19.fa" --BQSR
    "[dataset_path]/dataset_147605.dat" '
```

GATK2 v2.8.0: Print Reads

```
python [repo_tool_path]/gatk2/gatk2_wrapper.py --stdout
    "[dataset_path]/dataset_147610.dat" -d "-I"
    "[dataset_path]/dataset_147603.dat" "bam" "gatk_input" -d ""
    "[metadata_path]/metadata_16762.dat" "bam_index" "gatk_input" -p '
    java -jar "$GATK2_PATH/GenomeAnalysisTK.jar" -T "PrintReads" -o
    "[dataset_path]/dataset_147609.dat" $GATK2_SITE_OPTIONS --
    num_cpu_threads_per_data_thread ${GALAXY_SLOTS:-7} -R
    "[genome_path]/hg19.fa" --BQSR "[dataset_path]/dataset_147605.dat" -
    -disable_bam_indexing '
```

Box A4 - Short read data processing steps as executed by Galaxy (3). Continued from Box A3.

samtools v1.2: MPileup

```
samtools mpileup -f "[genome_path]/hg19.fa"
    "[dataset_path]/dataset_147609.dat" -s --output
    "[dataset_path]/dataset_147611.dat" 2>
    "[dataset_path]/dataset_147612.dat"
```

VarScan v2.3.5: InDels

```
perl [tool_path]/varscan/varscan_mpileup.pl "COMMAND::java -jar
    $JAVA_JAR_PATH/VarScan.v2.3.5.jar mpileup2indel"
    "INPUT::[dataset_path]/dataset_147611.dat"
    "OUTPUT::[dataset_path]/dataset_147613.dat"
    "LOG::[dataset_path]/dataset_147614.dat" "OPTION::--min-coverage 4"
    "OPTION::--min-reads2 2" "OPTION::--min-avg-qual 13" "OPTION::--min-
    var-freq 0.01" "OPTION::--min-freq-for-hom 0.75" "OPTION::--p-value
    0.95" "OPTION::--strand-filter 1" "OPTION::--output-vcf 1"
    "OPTION::--variants 1"
```

VarScan v2.3.5: SNPs

```
perl [tool_path]/varscan/varscan_mpileup.pl "COMMAND::java -jar
    $JAVA_JAR_PATH/VarScan.v2.3.5.jar mpileup2snp"
    "INPUT::[dataset_path]/dataset_147611.dat"
    "OUTPUT::[dataset_path]/dataset_147615.dat"
    "LOG::[dataset_path]/dataset_147616.dat" "OPTION::--min-coverage 4"
    "OPTION::--min-reads2 2" "OPTION::--min-avg-qual 13" "OPTION::--min-
    var-freq 0.01" "OPTION::--min-freq-for-hom 0.75" "OPTION::--p-value
    0.95" "OPTION::--strand-filter 1" "OPTION::--output-vcf 1"
    "OPTION::--variants 1"
```

VCFtools v0.1: Merge

```
vcf-sort [dataset_path]/dataset_147607.dat > 0.vcf.sorted ; bgzip
    0.vcf.sorted ; tabix -p vcf 0.vcf.sorted.gz ; vcf-sort
    [dataset_path]/dataset_147615.dat > 1.vcf.sorted ; bgzip
    1.vcf.sorted ; tabix -p vcf 1.vcf.sorted.gz ; vcf-sort
    [dataset_path]/dataset_147613.dat > 2.vcf.sorted ; bgzip
    2.vcf.sorted ; tabix -p vcf 2.vcf.sorted.gz ; vcf-merge
    0.vcf.sorted.gz 1.vcf.sorted.gz 2.vcf.sorted.gz >
    [dataset_path]/dataset_147617.dat
```

SNPeff v4.0

```
SNPEFF_JAR_PATH=[repo_tool_path]/snpEff/ SNPEFF_DATA_DIR=`grep
    '^data_dir' $SNPEFF_JAR_PATH/snpEff.config | sed
    's/.*data_dir.*[=:]//'`; eval "if [ ! -e $SNPEFF_DATA_DIR/hg19 ] ;
    then java -Xmx6G -jar $SNPEFF_JAR_PATH/snpEff.jar download -c
    $SNPEFF_JAR_PATH/snpEff.config hg19 ; fi"; java -Xmx6G -jar
    $SNPEFF_JAR_PATH/snpEff.jar eff -c $SNPEFF_JAR_PATH/snpEff.config -i
    vcf -o vcf -upDownStreamLen 5000 -spliceSiteSize 1 -lof -stats
    [dataset_path]/dataset_147619.dat -chr "chr" -noLog hg19
    [dataset_path]/dataset_147617.dat >
    [dataset_path]/dataset_147618.dat
```

Box A5 - Shell commands for plotting the genomic position vs. quality of called variants. Values are retrieved from the VCF 'POS' and 'QUAL' columns. Graph plots generated for every chromosome of a patient each are combined iteratively.

```
fire2="014 036 072 111 125 137 155"
fire3="020 090 213 281 344 375 406 428 566 586 598 624 638 708 750 796"

pair="start_vs_qual"
mkdir "$pair"
cd "$pair"
for pat in `echo $fire2 $fire3` ; do
    ~/project/gemini_plotter_v5.R \
      -d ../../gemini_dbs/all_merged.decomposed.normalized.vcf.db \
      -f $pat -x "start" -y "qual" \
      -G --no_local_titles --sample_global_titles \
      -w "qual != 'None' and qual >= 20 and qual < 10000" \
      --gt-filter "\"(gt_quals).(family_id = '"$pat"').(>=0).(any)\"" \
      --exonic-coloring;
    gm convert `ls *$pat*$pair*chr{1..9}.png` +append A.png
    gm convert `ls *$pat*$pair*chr{10..22}.png` +append B.png
    gm convert *$pat*$pair*chrX.png *$pat*$pair*chrY.png \
      *$pat*$pair*chrM.png +append C.png
    gm convert A.png B.png C.png +append "All_"$pat"_"$pair".png"
    if [ -f "All_VC_all_pats_all_chr_"$pair".png" ]; then
       gm convert "All_VC_all_pats_all_chr_"$pair".png" \
         "All_"$pat"_"$pair".png" -append \
         "All_VC_all_pats_all_chr_"$pair".png"
    else
       cp "All_"$pat"_"$pair".png" "All_VC_all_pats_all_chr_"$pair".png"
    fi
    rm "All_"$pat"_"$pair".png"
done
rm A.png B.png C.png
cd ..

pair="start_vs_qual(log)"
mkdir "$pair"
cd "$pair"
for pat in `echo $fire2 $fire3`; do
    ~/project/gemini_plotter_v5.R \
      -d ../../gemini_dbs/all_merged.decomposed.normalized.vcf.db \
      -f $pat -x "start" -y "qual" \
      -G -l "y" --no_local_titles --sample_global_titles \
      -w "qual != 'None' and qual >= 20 and qual < 10000" \
      --gt-filter "\"(gt_quals).(family_id = '"$pat"').(>=0).(any)\"" \
      --exonic-coloring;
    gm convert `ls *$pat*$pair*chr{1..9}.png` +append A.png
    gm convert `ls *$pat*$pair*chr{10..22}.png` +append B.png
    gm convert *$pat*$pair*chrX.png *$pat*$pair*chrY.png \
      *$pat*$pair*chrM.png +append C.png
    gm convert A.png B.png C.png +append "All_"$pat"_"$pair".png"
    if [ -f "All_VC_all_pats_all_chr_"$pair".png" ]; then
       gm convert "All_VC_all_pats_all_chr_"$pair".png" \
         "All_"$pat"_"$pair".png" -append \
         "All_VC_all_pats_all_chr_"$pair".png"
    else
       cp "All_"$pat"_"$pair".png" "All_VC_all_pats_all_chr_"$pair".png"
    fi
    rm "All_"$pat"_"$pair".png"
done
rm A.png B.png C.png
cd ..
```

Box A6 - Shell commands for plotting the genomic position vs. the variant caller-specific quality of called variants. Values are retrieved from the VCF 'POS' column and the 'GQ' field within the 'INFO' column (accessed by GEMINI as 'gt_quals'). Graph plots generated for every chromosome of a patient each are combined iteratively. As colors were used to separate variants from the particular calling tool, no exonic coloring has been used.

```
fire2="014 036 072 111 125 137 155"
fire3="020 090 213 281 344 375 406 428 566 586 598 624 638 708 750 796"

pair="start_vs_gt_quals"
mkdir "$pair"
cd "$pair"
for pat in `echo $fire2 $fire3` ; do
    ~/project/gemini_plotter_v5.R \
      -d ../../gemini_dbs/all_merged.decomposed.normalized.vcf.db \
      -f $pat -x "start" -y "gt_quals" \
      -G --no_local_titles --sample_global_titles \
      -w "qual != 'None' and qual >= 20 and qual < 10000"
      --gt-filter "\"(gt_quals).(family_id = '"$pat"').(>=0).(any)\"";
    gm convert `ls *$pat*$pair*chr{1..9}.png` +append A.png
    gm convert `ls *$pat*$pair*chr{10..22}.png` +append B.png
    gm convert *$pat*$pair*chrX.png *$pat*$pair*chrY.png \
      *$pat*$pair*chrM.png +append C.png
    gm convert A.png B.png C.png +append "All_"$pat"_"$pair".png"
    if [ -f "All_VC_all_pats_all_chr_"$pair".png" ]; then
      gm convert "All_VC_all_pats_all_chr_"$pair".png" \
        "All_"$pat"_"$pair".png" -append \
        "All_VC_all_pats_all_chr_"$pair".png"
    else
      cp "All_"$pat"_"$pair".png" "All_VC_all_pats_all_chr_"$pair".png"
    fi
    rm "All_"$pat"_"$pair".png"
done
rm A.png B.png C.png
cd ..

pair="start_vs_gt_quals(log)"
mkdir "$pair"
cd "$pair"
for pat in `echo $fire2 $fire3`; do
    ~/project/gemini_plotter_v5.R \
      -d ../../gemini_dbs/all_merged.decomposed.normalized.vcf.db \
      -f $pat -x "start" -y "gt_quals" \
      -G -l "y" --no_local_titles --sample_global_titles \
      -w "qual != 'None' and qual >= 20 and qual < 10000" \
      --gt-filter "\"(gt_quals).(family_id = '"$pat"').(>=0).(any)\"";
    gm convert `ls *$pat*$pair*chr{1..9}.png` +append A.png
    gm convert `ls *$pat*$pair*chr{10..22}.png` +append B.png
    gm convert *$pat*$pair*chrX.png *$pat*$pair*chrY.png \
      *$pat*$pair*chrM.png +append C.png
    gm convert A.png B.png C.png +append "All_"$pat"_"$pair".png"
    if [ -f "All_VC_all_pats_all_chr_"$pair".png" ]; then
      gm convert "All_VC_all_pats_all_chr_"$pair".png" \
        "All_"$pat"_"$pair".png" -append \
        "All_VC_all_pats_all_chr_"$pair".png"
    else
      cp "All_"$pat"_"$pair".png" "All_VC_all_pats_all_chr_"$pair".png"
    fi
    rm "All_"$pat"_"$pair".png"
done
rm A.png B.png C.png
cd ..
```

Box A7 - Shell commands for plotting the depth of coverage vs. the variant caller-specific quality of called variants (1). Values are retrieved from the 'DP' and the 'GQ' field within the VCF 'INFO' column (accessed by GEMINI as 'gt_depth' and 'gt_quals'). Graph plots generated for every chromosome of a patient each are combined iteratively. As colors were used to separate variants from the particular calling tool, no exonic coloring has been used.

```
fire2="014 036 072 111 125 137 155"
fire3="020 090 213 281 344 375 406 428 566 586 598 624 638 708 750 796"

pair="gt_depths_vs_gt_quals"
mkdir "$pair"
cd "$pair"

for pat in `echo $fire2 $fire3` ; do
    ~/project/gemini_plotter_v5.R \
      -d ../../gemini_dbs/all_merged.decomposed.normalized.vcf.db \
      -f $pat -x "gt_depths" -y "gt_quals" \
      -G --no_local_titles --sample_global_titles \
      -w "qual != 'None' and qual >= 20 and qual < 10000" \
      --gt-filter "\"(gt_quals).(family_id = '"$pat"').(>=0).(any)\"";
    gm convert `ls *$pat*$pair*chr{1..9}.png` +append A.png
    gm convert `ls *$pat*$pair*chr{10..22}.png` +append B.png
    gm convert *$pat*$pair*chrX.png *$pat*$pair*chrY.png \
      *$pat*$pair*chrM.png +append C.png
    gm convert A.png B.png C.png +append "All_"$pat"_"$pair".png"
    if [ -f "All_VC_all_pats_all_chr_"$pair".png" ]; then
      gm convert "All_VC_all_pats_all_chr_"$pair".png" \
        "All_"$pat"_"$pair".png" -append \
        "All_VC_all_pats_all_chr_"$pair".png"
    else
      cp "All_"$pat"_"$pair".png" "All_VC_all_pats_all_chr_"$pair".png"
    fi
    rm "All_"$pat"_"$pair".png"
done
rm A.png B.png C.png
cd ..

pair="gt_depths(log)_vs_gt_quals"
mkdir "$pair"
cd "$pair"
for pat in `echo $fire2 $fire3` ; do
    ~/project/gemini_plotter_v5.R \
      -d ../../gemini_dbs/all_merged.decomposed.normalized.vcf.db \
      -f $pat -x "gt_depths" -y "gt_quals" \
      -G -l "x" --no_local_titles --sample_global_titles \
      -w "qual != 'None' and qual >= 20 and qual < 10000" \
      --gt-filter "\"(gt_quals).(family_id = '"$pat"').(>=0).(any)\"";
    gm convert `ls *$pat*$pair*chr{1..9}.png` +append A.png
    gm convert `ls *$pat*$pair*chr{10..22}.png` +append B.png
    gm convert *$pat*$pair*chrX.png *$pat*$pair*chrY.png \
      *$pat*$pair*chrM.png +append C.png
    gm convert A.png B.png C.png +append "All_"$pat"_"$pair".png"
    if [ -f "All_VC_all_pats_all_chr_"$pair".png" ]; then
      gm convert "All_VC_all_pats_all_chr_"$pair".png" \
        "All_"$pat"_"$pair".png" -append \
        "All_VC_all_pats_all_chr_"$pair".png"
    else
      cp "All_"$pat"_"$pair".png" "All_VC_all_pats_all_chr_"$pair".png"
    fi
    rm "All_"$pat"_"$pair".png"
done
rm A.png B.png C.png
cd ..
```

Box A8 - Shell commands for plotting the depth of coverage vs. the variant caller-specific quality of called variants (2). Continued from Box A7.

```
fire2="014 036 072 111 125 137 155"
fire3="020 090 213 281 344 375 406 428 566 586 598 624 638 708 750 796"

pair="gt_depths(log)_vs_gt_quals(log)"
mkdir "$pair"
cd "$pair"
for pat in `echo $fire2 $fire3` ; do
    ~/project/gemini_plotter_v5.R
      -d ../../gemini_dbs/all_merged.decomposed.normalized.vcf.db \
      -f $pat -x "gt_depths" -y "gt_quals" \
      -G -l "xy" --no_local_titles --sample_global_titles \
      -w "qual != 'None' and qual >= 20 and qual < 10000" \
      --gt-filter "\"(gt_quals).(family_id = '"$pat"').(>=0).(any)\"";
    gm convert `ls *$pat*$pair*chr{1..9}.png` +append A.png
    gm convert `ls *$pat*$pair*chr{10..22}.png` +append B.png
    gm convert *$pat*$pair*chrX.png *$pat*$pair*chrY.png \
      *$pat*$pair*chrM.png +append C.png
    gm convert A.png B.png C.png +append "All_"$pat"_"$pair".png"
    if [ -f "All_VC_all_pats_all_chr_"$pair".png" ]; then
      gm convert "All_VC_all_pats_all_chr_"$pair".png" \
        "All_"$pat"_"$pair".png" -append \
        "All_VC_all_pats_all_chr_"$pair".png"
    else
      cp "All_"$pat"_"$pair".png" "All_VC_all_pats_all_chr_"$pair".png"
    fi
    rm "All_"$pat"_"$pair".png"
done
rm A.png B.png C.png
cd ..
```

8. Danksagung

Eine Promotion schreibt sich ganz offensichtlich nicht von selbst – und nur selten geht ein Doktorand vollständig alleine durch diese Zeit. In allen anderen Fällen gebührt den Unterstützern Dank.

Im Rahmen in dieser Thesis hat mich eine Reihe von Leuten unterstützt, fachlich wie persönlich. Am Anfang der Liste steht selbstverständlich Prof. Ulrich Mansmann, der nicht nur als Doktorvater und Themengeber den akademischen Rahmen dieser Arbeit hergestellt hatte, sondern auch über sechs Jahre und zwei Projekte eine großartige Stelle für mich geschaffen und hochgehalten hat. Ich habe in dieser Zeit größte Freiheiten genießen und maximal selbstständig arbeiten dürfen. Ein nachhaltiges NGS-IT-System ist die Folge.

Persönlich an erster Stelle aber steht meine Frau Wiebke Schaaf, die nicht nur tapfer mit durchgehalten und mich unterstützt hat, sondern zum Schluss auch das beste Argument für einen Schlusspunkt unter diese Arbeit auf die Welt gebracht hat. Danke für die Motivation – auch Dir, Oliver!

Auch am IBE zeichnen sich eine lange Reihe von Freunden und Kollegen dafür verantwortlich, für eine großartige Atmosphäre, eine tolle Zeit und denkwürdige Augenblicke gesorgt zu haben. Auf eine Reihenfolge möchte ich mich gar nicht festlegen, daher nachfolgend die Wichtigsten grob den Flur entlang: Tobias Schleinkofer, Ronja Woltersdorf, Anna Rieger, Monika Jelizarow, Vindi Jurinovic, David Reinhardt, Heidi Seibold, Verena Hoffmann, Michael Lauseker, Markus Pfirrmann und Miriam Rottmann. Für großartige Unterstützung in Technik und Organisation Klaus Rüschstroer, Nikolaus von Bomhard, Wolfgang Brummer und unsere Sekretärinnen. Ein großer Dank geht an Guokun Zhang, Sandra Fischer und Fabian Grandke für herausragende Teamarbeit. Zusätzlich ein riesengroßes „Danke!“ an Aarif Mohamed Nazeer Batcha: es war mir eine große Ehre, mit Dir zusammenarbeiten zu können.

Durch die Projektarbeit war es letztlich unmöglich, an der Klinikums-IT vorbei zu kommen – und es wäre auch mehr als schade gewesen. Danke an alle MIT’ler, die mich unterstützt haben - ohne euch hätte das alles nicht funktioniert. Hervorheben möchte ich Sammy Simba, Simon Leutner und Frank Hülle. Die wohl wichtigste Rolle hat allerdings Gregor Pickert gespielt: danke, dass Du Dich erst von unserem Projekt, und schließlich von meiner Arbeit hast überzeugen lassen. Danke für offene Ohren.

Auch aus anderen Institutionen haben Leute zum Verlauf beigetragen: Prof. Andreas Jung (Pathologie), ein großer Teil des Labors für Leukämiediagnostik der Medizinischen Klinik III (danke speziell an Maja Rothenberg-Thurley), die AG Blum des Genzentrums LMU, aus letzterer besonders Alexander Graf. Er hat zu verantworten, dass ich mit „Galaxy“ in der meiner Meinung nach großartigsten Community der internationalen Bioinformatik gelandet bin. Die Liste wäre viel zu lang, daher stellvertretend, aber auch persönlich danke an Björn Grüning für Aufnahme und Unterstützung über all die Jahre.

Der Firma Merck sei an dieser Stelle für die Finanzierung der Datengenerierung gedankt.

Juliane Fluck, Martin Hofmann-Apitius und Jens Dörpinghaus danke ich herzlich für Geduld, Verständnis und Unterstützung bei der Ausarbeitung der Thesis. Auch hier: ohne euch wäre das nicht gegangen.

Der Prüfungskommission, bestehend aus Prof. Ulrich Mansmann, Prof. Regina Fluhrer, Prof. Roland Kappler und PD Andreas Herbst danke ich für die erfolgreich abgehaltene Prüfung sowie die Begutachtung der schriftlichen Ausarbeitung. Für Letzteres bedanke ich mich ebenfalls bei Prof. Volker Heun.

Gedenken möchte ich Klaus Rüschstroer, Jacek Puchałka, Michael Specht, Friedel Pasman und Lother Bresges, die das Ende der Thesis leider nicht mehr erleben konnten.

Abschließend danke ich Frank Birkenhauer (der den Titel schon vor einer Ewigkeit vorhergesagt hat) für das Vertrauen und meiner Familie, speziell meinen Eltern, dafür, dass sie den Grundstein gelegt hat für das Erreichte. Dass das alles nicht ganz von der Stange, um nicht zu sagen „eigenartig“ gelaufen ist, ist sicher auch euch geschuldet – und das ist gut so. Denn „eigen“ (selbstständig) und „artig“ (korrekt) habe ich von euch gelernt. Danke.