Abstract

Rare disorders are defined as life-threatening or chronically debilitating diseases affecting less than 5 per 10,000 people, but collectively they are estimated to affect 5-8% of the world population. The standard of care for rare disorders is inferior to the standard of care for common disorders in two ways. First, the speed and effectiveness of the diagnostic process is lacking, with an average time to diagnosis of 5 to 30 years and with 40% of the patients being initially misdiagnosed. Second, even after a correct diagnosis, effective therapies are rarely available. Since most rare disorders have a genetic cause, the identification of the causative genes is the first step in both delivering a correct molecular diagnosis and in the development of effective therapies. The advent of next generation sequencing has greatly simplified gene identification, but the analysis of NGS data still provides many challenges.

This PhD thesis had two main aims. First, to generate an automated bioinformatics pipeline for the analysis of NGS data, which brings together published, state-of-the-art tools with novel algorithms developed by me. Second, to identify novel genes which cause premature aging syndromes. The pipeline was used for gene identification in a wide spectrum of rare disorders in collaboration with other members of the laboratory and with other groups. In total, we identified 19 novel disease-causing genes, 12 novel candidate disease-causing genes, 3 novel genes causing specific zebrafish phenotypes, and 11 known disease-causing genes. The novel bioinformatics tools I developed include ExomeIBD, a protocol for the identification of *de novo* mutations, a protocol for the detection of CNVs, a filtering strategy for WES data based on protein-protein interactions, and a mapping tool to locate causative mutations in zebrafish WES.

ExomeIBD is an identity-by-descent mapping tool, which can achieve similar results to standard array-based techniques, using only data derived from WES. I used ExomeIBD to show that homozygous regions deriving form recent consanguineous unions are less affected by selection pressure, when compared to homozygous regions deriving from background inbreeding in the population. Additionally, ExomeIBD was used as a filter for WES data, leading to the identification of *WNT1* as a

causative gene for OI. The protocol for *de novo* analysis that I developed was used to identify causative genes for Hallermann-Streiff syndrome and Kabuki syndrome. With a combination of CNV analysis and breakpoint analysis in WES data, I was able to identify *CRIM1* as the causative gene for MACOM syndrome. The protein-protein interaction filter that I developed was used to identify two *LIG4* mutations, which were missed in standard analysis in a patient with primary microcephalic dwarfism. I further developed a mapping tool for zebrafish WES experiments, which we used as a filter to identify a mutation in *atp1b1a* in mutant fish showing a skin-aggregation phenotype and a mutation in *wwox* in mutant fish showing a fainting-dwarf phenotype.

In the WES data of a patient affected by progeria with long survival, I identified a variation in *LONP1*, which is highly enriched in a cohort of premature aging patients. LONP1 is a mitochondrial protease, which is also involved in the replication of mtDNA. Western blot analysis of LONP1 in fibroblasts derived from two different patients showed reduced protein levels of LONP1. In a collaborative effort, we could show that fibroblasts from the two patients have an 85% and 15% decrease in the amount of mtDNA, respectively. These results indicate that the progeria with long survival might be a mtDNA-depletion syndrome.

Additionally, using WES, I identified a *de novo* mutation in *ANO6* in a patient affected by neonatal progeria syndrome. ANO6 is a scramblase, which can eliminate the lipid gradients between the two leaflets of the plasma membrane. In patient fibroblasts, I could show that the basal activation of the MEK/ERK signaling pathway is stronger than in control fibroblasts, as observed in western blot analysis of p-ERK, p-MEK and p-cRAF. These results, together with the known role of MEK/ERK dysregulation in the aging process, led to the hypothesis that gain-of-function mutations in ANO6 cause alterations in the lipid composition of the plasma membrane, which in turn lead to the activation of the MEK/ERK pathway.

These successful gene-identification studies allow for diagnostic testing and open up the opportunity to perform detailed molecular analysis in order to understand the molecular pathogenesis of these disorders.

Finally, using a novel filtering strategy, I show that NGS data can be used

to measure the amount of somatic mutations in a tissue and that patients with genomic instability syndromes show an increased amount of somatic mutations, when compared to patients with other syndromes. This novel method for measuring somatic mutations will be valuable in the study of the pathogenesis of accelerated aging syndromes and of the basic mechanisms underlying physiological aging.

Zusammenfassung

Als seltene Krankheiten werden lebensbedrohliche oder chronisch schwächende Erkrankungen bezeichnet, deren Inzidenz weniger als 5 von 10.000 Menschen beträgt. Zusammenfassend betrachtet leiden jedoch 5-8% der Weltbevölkerung an einer seltenen Erkrankung. Der Versorgungsstandard bei seltenen Erkrankungen ist wesentlich schlechter als bei häufigen Erkrankungen. Zum einen mangelt es dem diagnostischen Prozess an Schnelligkeit und Effektivität, wobei eine Diagnose durchschnittlich 5 bis 30 Jahre in Anspruch nehmen kann und 40% der Betroffenen zunächst fehldiagnostiziert werden. Des Weiteren stehen selbst nach einer korrekten Diagnose nur selten effektive Therapien zur Verfügung. Die meisten seltenen Erkrankungen beruhen auf einer genetischen Ursache. Deshalb ist die Identifizierung von krankheitsverursachenden Mutationen der erste Schritt, um eine korrekte molekulare Diagnose zu stellen und die Entwicklung von Therapien zu ermöglichen. Die Einführung der next generation sequencing (NGS) Technik hat die Identifizierung von ursächlichen Genvariationen maßgeblich vereinfacht. Die Analyse der NGS-Daten stellt jedoch viele Herausforderungen dar.

Die vorliegende Doktorarbeit hatte zwei Aufgaben zum Ziel: Erstens sollte eine automatische Bioinformatik-Pipeline zur Analyse von NGS-Daten entwickelt werden, die bereits publizierte, hochmoderne Arbeitsprozesse mit neuen, von mir entwickelten Algorithmen verbindet. Zweitens sollten neue Krankheitsgene identifiziert werden, die Syndrome mit vorzeitiger Alterung verursachen. Die Pipeline wurde für ein breites Spektrum von seltenen Erkrankungen in Zusammenarbeit mit den Mitarbeitern der Arbeitsgruppe und in Kollaboration mit anderen Gruppen verwendet. Insgesamt konnten wir 19 neue krankheitsverursachende Gene identifizieren, 12 neue Kandidatengene, 3 neue Gene, die bestimmte Phänotypen im Zebrafisch verursachen, sowie 11 bekannte Krankheitsgene. Die von mir neu entwickelte bioinformatische Methode beinhaltet ExomeIBD, ein Protokoll zur Identifizierung von *de novo* Mutationen, ein Protokoll zur Analyse von *copy number variations* (CNVs), eine proteininteraktionsbasierte Filterstrategie für *whole-exome sequencing* (WES) und eine Methode zur Kartierung

von WES-Daten aus dem Zebrafisch.

ExomeIBD ist ein *identity–by–descent* mapping tool, welches in der Lage ist, herkömmliche Homozygotie-Kartierungsexperimente nur unter Verwendung der Daten vom WES zu reproduzieren. Damit konnte ich zeigen, dass homozygote Regionen von phylogenetisch jüngeren konsanguinen Verbindungen weniger vom Selektionsdruck betroffen sind als homozygote Regionen, welche aus *background inbreeding* in der Population stammen. ExomeIBD wurde weiterhin als Filter für WES-Daten benutzt, um *WNT1* als neues OI-verursachender Gen zu identifizieren. Das Protokoll für die Analyse von *de novo* Mutationen konnte erfolgreich eingesetzt werden, um kausale Mutationen in bisher nicht assoziierten Genen für das Hallermann-Streiff- und das Kabuki-Syndrom zu identifizieren. Durch die Kombination der Analyse von CNVs und Bruchpunkten aus WES-Daten gelang es mir, *CRIM1* als kausales Gen für das MACOM-Syndrom zu identifizieren. Die von mir entwickelte proteininteraktionsbasierte Filterstrategie wurde in einem Patienten mit primärer Mikrozephalie und Kleinwuchs erfolgreich eingesetzt, um Mutationen in *LIG4* zu identifizieren, welche bei Anwendung von Standardanalysemethoden übersehen worden wären. Weiterhin entwickelte ich eine Methode zur Kartierung von WES-Daten aus dem Zebrafisch, die als Filter genutzt wurde, um eine Mutation in *atp1b1a* in einem mutierten Zebrafisch mit einem Hautaggregationsphänotyp zu identifizieren sowie eine Mutation in *wwox* in einem mutierten Zebrafisch, der durch Kleinwuchs und Ohnmachtsanfälle gekennzeichnet ist.

In den WES-Daten eines Progeriepatienten mit langer Lebenserwartung habe ich eine Variante in *LONP1* identifiziert, welche in einer Kohorte von Progerie-Patienten vermehrt auftritt. Bei LONP1 handelt es sich um eine mitochondriale Protease, die an der mtDNA-Replikation beteiligt ist. Western blot-Analysen von LONP1 aus Fibroblasten zweier Patienten zeigten eine reduzierte Expression von LONP1. In Kollaboration mit der Arbeitsgruppe um Prof. Rudolf Wiesner konnten wir zeigen, dass Patientenfibroblasten 85% weniger mtDNA beinhalten, was darauf hindeuten könnte, dass es sich bei dieser Progerie um ein mtDNA-Deletionssyndrom handelt.

Mittels WES konnte ich weiterhin eine *de novo* Mutation in *ANO6* in einem Patienten mit neonatalem Progerie-Syndrom identifizieren. ANO6 ist eine

Scramblase, welche den Lipidgradienten zwischen den beiden Plasmamembranteilen aufheben kann. Anhand von Western blot-Analysen von p-ERK, p-MEK and p-cRAF konnte ich zeigen, dass die Aktivierung des MEK/ERK-Signalwegs in Patientenfibroblasten wesentlich stärker ausgeprägt ist als in Kontrollfibroblasten. Diese Ergebnisse zusammen mit dem Wissen, dass die MEK/ERK-Dysregulation eine wichtige Rolle im Alterungsprozess spielt, lassen darauf schließen, dass *gain-of-function* Mutationen in ANO6 eine Veränderung der Lipidzusammensetzung der Plasmamembran hervorrufen, welche wiederum zur Aktivierung des MEK/ERK-Signalwegs führt.

Diese erfolgreichen Genidentifizierungstudien ermöglichen eine diagnostische Testung und eröffnen die Möglichkeit, detaillierte molekulare Analysen durchzuführen, um die molekulare Pathogenese dieser Krankheiten verstehen zu können.

Abschließend konnte ich mithilfe einer neuen Filterstrategie zeigen, dass NGS-Daten ebenfalls dazu verwendet werden können, um die Menge an somatischen Mutationen in einem Gewebe zu bestimmen und dass Patienten mit Syndromen, bei denen genomische Instabilität vorliegt, im Vergleich zu Patienten mit anderen Syndromen wesentlich mehr somatische Mutationen aufweisen. Diese neue Methode zur Bestimmung somatischer Mutationen soll bei der Erforschung der Pathogenese von Syndromen mit frühzeitiger Alterung helfen, kann aber auch zur Klärung grundlegender Mechanismen beitragen, die eine Rolle beim Prozess des physiologischen Alterns spielen.