Aus der Klinik für Pädiatrie m.S. Neurologie
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Erstellung eines computergestützten Verfahrens zur Suche nach Kandidatengenen für rezessiv vererbte Krankheiten in konsanguinen Familien

zur Erlangung des akademischen Grades

Doctor rerum medicarum (Dr. rer. medic.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Dominik Seelow
aus Berlin

Gutachter: 1. Prof. Dr. M. Schülke-Gerstenfeld

2. Prof. Dr. K. Sperling

3. Prof. Dr. A. Reis

Datum der Promotion: 4.12.2009

Inhaltsverzeichnis

Abstract

Diese Publikationspromotion umfasst sechs publizierte Arbeiten, die sich mit der Suche nach krankheitsverursachenden Genmutationen insbesondere in konsanguinen Familien befassen. Hierfür habe ich eine Reihe von Computerprogrammen entwickelt, die es den Wissenschaftlern ermöglichen, schneller und einfacher als bisher in Frage kommende chromosomale Regionen zu identifizieren, vielversprechende Kandidatengene auszusuchen und gefundene Mutationen auf ihr krankheitsverursachendes Potential zu untersuchen.

Dazu habe ich eine Datenbank erstellt, in der zahlreiche genspezifische Daten integriert sind. Diese Datenbank wird von sämtlichen Applikationen genutzt. Mit *HomozygosityMapper* (http://www.homozygositymapper.org) habe ich ein neues Verfahren zur Homozygotiekartierung implementiert, das teilweise um mehrere Größenordnungen schneller arbeitet als klassische Verfahren. *GeneDistiller* (http://www.genedistiller.org) ist eine Suchmaschine für Kandidatengene, die umfangreiche genetische Daten übersichtlich zusammenfasst und die Forscher intuitiv und interaktiv bei der Auswahl von Kandidatengenen unterstützt. *HomozygosityMapper* und *GeneDistiller* sind miteinander vernetzt, so dass ausgewählte Gene auf Homozygotie untersucht und aussichtsreiche Gene in homozygoten Regionen bestimmt werden können.

Diese Applikationen werden erweitert durch *AssociationDB* (http://compbio.charite.de/genetik/AssociationDB/), eine Datenbank für Assoziationsstudien, sowie durch *FragIdent* (http://compbio.charite.de/genetik/FragIdent/), ein Programm, welches cDNA Sequenzen Genen zuordnen kann und mit weiteren Informationen wie z.B. Proteindomänen ergänzt. Zusätzlich zu den hier genannten Computerprogrammen konnte in einer Studie gezeigt werden, dass das Verfahren der Homozygotiekartierung auch für nicht-konsanguine Familien mit rezessiven Erkrankungen eingesetzt werden kann.

Einleitung

Konsanguine Familienstrukturen begünstigen das Auftreten rezessiv vererbter Krankheiten. Trägt ein Vorfahre (Founder) eine krankheitsverursachende Variante eines Gens (=Krankheitsallel), kann diese heterozygot vererbt werden, ohne dass die Träger(innen) selbst erkranken. Wenn allerdings zwei Nachkommen dieser Founderperson mit jeweils einem Krankheitsallel Kinder bekommen, ist es möglich, dass diese von jedem Elternteil die krankheitsverursachende Variante erben, für diese also homozygot (autozygot) sind und erkranken. Die üblichen Verfahren in der Homozygotiekartierung basieren auf der Mehrpunkt-Kopplungsanalyse (siehe Methodik) und wurden ursprünglich für Mikrosatelliten entwickelt. Sie sind deshalb für relativ wenige Marker mit einer hohen Informativität ausgelegt. Bei den heute verwendeten Genotypisierungs-Chips mit SNPs *(single nucleotide polymorphisms*, Einzelnukleotidpolymorphismen) werden jedoch sehr viele Marker (bis zu ca. 100.000 pro Chromosom) mit einer sehr geringen Informativität analysiert. Die klassische Mehrpunktanalyse ist somit hierfür nur eingeschränkt einsetzbar. Neben ihrer Anfälligkeit für Genotypisierungsfehler ist vor allem der zum Teil enorme Zeitaufwand unbefriedigend.

Wenn eine gemeinsam mit der Krankheit vererbte Region gefunden wird, muss in dieser die ursächliche Genmutation identifiziert werden. Da insbesondere in Studien mit nur wenigen Betroffenen diese Regionen sehr groß sein und mehrere hundert Gene enthalten können, müssen Wissenschaftler zuerst die wahrscheinlichsten Krankheitsgene ermitteln und diese dann gezielt durch Sequenzierung auf Abweichungen der DNA-Sequenz zwischen Betroffenen und Kontrollpersonen, z.B. gesunden Familienmitgliedern, hin untersuchen. Die Auswahl der Gene geschieht in der Regel manuell durch das Studium von Internet-Datenquellen sowie Publikationen zu allen in dieser Region befindlichen Genen; häufig werden auch Computerprogramme eingesetzt, die auf Grundlage definierter Kriterien Kandidatengene vorschlagen. Während der manuelle Ansatz eine hohe Flexibilität ermöglicht, ist er jedoch extrem zeitaufwendig. Die bisherigen Softwarelösungen sind hingegen 'black boxes', die das Hintergrundwissen der Wissenschaftler in der Regel nicht ausreichend integrieren.

Wird eine Mutation in einem Kandidatengen gefunden, so muss diese als krankheitsverursachend validiert werden. Die Generierung von Tiermodellen ist aus Zeit- und Kostengründen in der Regel keine Option. Stattdessen wird meist nachgewiesen, dass die Mutation in einer ausreichend großen Zahl gesunder Kontrollpersonen aus derselben Population nicht vorkommt. Die dazu erforderlichen Sequenzanalysen mehrerer hundert Proben sind sowohl zeit- als auch kostenaufwendig. Computerprogramme zur Vorhersage des Krankheitspotentials einer Mutation existieren zwar, fokussieren sich in der Regel aber auf einen einzelnen Aspekt, z.B. lediglich auf Änderungen der *splice sites* oder die Konservierung von Aminosäuren.

Zielstellung

Ziel dieser Arbeit war es, ein bioinformatisches Verfahren zu entwickeln, das deutlich schneller als die bisherigen Lösungen homozygote Genregionen in Patienten aus konsanguinen Familien detektieren und nach ihrer vermuteten Relevanz ordnen könnte. Darüber hinaus sollten Informationen zu den in den gefundenen Intervallen befindlichen Genen integriert werden, so dass die Wissenschaftler in die Lage versetzt werden, direkt die für die Erkrankung in Frage kommenden Kandidatengene zu bestimmen.

Methodik

Datenintegration

Zu den kommerziell erhältlichen SNP-Chips für die Homozygotiekartierung sind Annotationsdateien erhältlich, die Details zu den verwendeten SNP-Markern erhalten. Um eine bessere Qualität der Daten zu garantieren, wurden zusätzlich sämtliche derzeit bekannten SNPs aus dbSNP (1) in eine lokale Datenbank für *HomozygosityMapper* (2) übernommen. Dabei wurden mehrfach annotierte SNPs entsprechend gekennzeichnet, um diese von Analysen ausschließen zu können; die Verwendung der Positionsdaten aus dbSNP garantiert außerdem die gleiche Datenbasis für die SNPs auf verschiedenen Chips. Zusätzlich wurden die Genotypfrequenzen des HapMap Projekts (3) für mehr als 4 Millionen SNPs aufgenommen, um so fundierte Angaben über die Frequenz heterozgyoter Genotypen zu erhalten. Da von den mehr als 14 Millionen bekannten SNPs nur ca. 2 Millionen derzeit auf Genotypisierungs-Chips eingesetzt werden, wurde ein 'materialised view' als Datenbanktabelle erstellt, der nur die eingesetzten SNPs umfasst und Abfragen somit erheblich beschleunigt.
Für die Ermittlung aussichtsreicher Kandidatengene wurden zahlreiche im Internet verfügbaren Datenquellen integriert; eine Liste findet sich in der Publikation zu *GeneDistiller* (4). Darüber hinaus ist auch die *Human Phenotype Ontology* (5) Teil der Datenbank und kann ebenfalls zur Bestimmung von Kandidatengenen herangezogen werden.
Zur Vorhersage des Krankheitspotentials einer Sequenzveränderung wurden weitere Daten in unsere lokale Datenbank aufgenommen; hier sind beispielsweise funktionelle Proteinmotive aus Swiss-Prot (6) zu nennen. Die Datenbankschemata werden auf den Homepages der einzelnen Applikationen dargestellt.

Homozygotiekartierung

Bisher wird hierfür zumeist eine Mehrpunkt-Kopplungsanalyse eingesetzt, dabei kommen im Wesentlichen drei verschiedene Methoden zur Anwendung (7):

1. *Lander-Green-Algorithmus*
   Dieser skaliert linear mit der Anzahl der Marker, aber exponentiell mit der Anzahl der Personen (Beschränkung auf etwa 20 Individuen).

2. *Elston-Stewart-Algorithmus*
   Dieser skaliert linear mit der Anzahl der Personen, aber exponentiell mit der Anzahl der Marker und ist deshalb auf maximal etwa 8 Marker beschränkt. Er wird außerdem durch fehlende Genotypen stark verlangsamt.

3. Näherungsverfahren mit einem *Markov Chain / Monte Carlo Algorithmus.*.
   Diese Verfahren generieren zwar keine exakten LOD Scores, sind aber dafür - zumindest theoretisch - robuster gegen fehlende Daten als die beiden anderen Algorithmen und skalieren besser.

Alle oben genannten Verfahren werden durch konsanguine Familienstrukturen rechenaufwendiger. Die prominentesten Computerprogramme zur Mehrpunkt-Analyse sind

Allegro (8), GENEHUNTER und Derivate (9-11) (alle nutzen den Lander-Green-Algorithmus) sowie Simwalk2 (7) (Markov Chain / Monte Carlo).
Da es sich bei der Kopplungsanalyse um ein statistisches Verfahren handelt, ist es erforderlich, möglichst viele Betroffene sowie ihre gesunden Verwandten zu genotypisieren, um die Erbgänge nachvollziehen zu können. Im Falle konsanguiner Familien und erwarteter Autozygotie kann man jedoch auf die gesunden Familienangehörigen verzichten (12), da man hier lediglich nach homozgyoten Genregionen suchen muss. Während so die Probensammlung vereinfacht und der experimentelle Aufwand sowie die Kosten gesenkt werden, steigt der Bedarf an Rechenleistung bei Verwendung der konventionellen Computerprogramme enorm an, da für alle nicht typisierten Personen in einem Stammbaum Genotypen bzw. Haplotypen antizipiert werden müssen. Insbesondere die Berechnung von Haplotypen ist durch die geringere Informativität von SNPs gegenüber Mikrosatelliten sowie ihre große Zahl bei der Verwendung von Chips extrem rechenintensiv (13) und fehleranfällig (14).
Trotz der beschriebenen Probleme der bisherigen Lösungen werden diese weiterhin allgemein genutzt und durch den Kauf zusätzlicher (leistungsfähigerer) Computer oder das Weglassen einer großen Zahl genetischer Marker in der initialen Analyse kompensiert. Alternativ besteht auch die Möglichkeit, homozygote Regionen ohne Kopplungsanalyse durch eine einfache Zählung aufeinander folgender homozygoter Genotypen zu ermitteln. Derartige Lösungen (15;16) haben sich bisher jedoch nicht durchsetzen können.

HomozygosityMapper (2;2) liegt eine SNP-Datenbank zugrunde. In einem ersten Schritt werden alle Genotypen eines Projekts eingelesen, die auf einem dort eindeutig annotierten SNP basieren, wodurch zweifelhafte SNPs ignoriert werden. Nach dem Einlesen werden automatisch homozygote Blöcke in jeder Person detektiert, einzelne heterozygote Genotypen inmitten eines homozygoten Blocks werden dabei als Genotypisierungsfehler angesehen und beeinträchtigen die Erkennung somit nicht.
Die so gewonnenen Daten können beliebig oft analysiert werden. Hierbei können beliebig viele Personen aus einer oder mehreren Familien als Betroffene bzw. gesunde Kontrollen ausgewählt werden. *HomozygosityMapper* durchläuft nun das gesamte Genom und addiert die Längen der homozygoten Blöcke jedes Betroffenen an jeder Markerposition. Um eine Aufblähung der Werte durch einzelne sehr lange Blöcke zu vermeiden, wird die maximal addierte Länge begrenzt, dabei werden für die einzelnen Chips optimierte Werte verwendet. Diese lassen sich bei Bedarf, z.B. bei sehr entfernter Konsanguinität und somit der Erwartung relativ kurzer homozygoter Blöcke am Krankheitslocus, durch besser passende Einstellungen überschreiben.
Nach erfolgter Analyse wird ein genomweiter Homozygotie-Score grafisch angezeigt und vielversprechende Regionen werden optisch hervorgehoben und können wahlweise durch schrittweises Hineinzoomen oder durch direkte Hyperlinks näher betrachtet werden. Ebenso können die einzelnen zugrundeliegenden Genotypen visualisiert werden, wodurch eine optische Kontrolle der Regionen sowie einzelner Personen ermöglicht wird.
HomozygosityMapper ignoriert gesunde Kontrollen bei der Datenanalyse, um falsch negative Ergebnisse zu vermeiden, zeigt deren Genotypen aber mit an. Ebenso werden beide homozygoten Genotypen gleich betrachtet, da in verschiedenen Familien – auch bei einer Mutation im gleichen Gen - nicht zwangsläufig ein gemeinsamer Krankheitshaplotyp zu erwarten ist. In der Anzeige der Genotypen werden aber die Originalgenotypen dargestellt, so dass zwischen beiden homozygoten Genotypen unterschieden werden kann. Darüber hinaus

werden hier auch die Genotypen der Kontrollen dargestellt, so dass ein manueller Ausschluss von Regionen durch diese Informationen leicht möglich ist.
Die Grenzen einer Region werden in der Genotypenansicht deutlich dargestellt und können durch Mausklicks verändert werden, um beispielsweise Teilregionen auf Grundlage nicht betroffener Familienmitglieder auszuschließen.

Suche nach Kandidatengenen

Um die Suche nach Kandidatengenen innerhalb einer genomischen Region so effizient wie möglich zu gestalten, erschien es uns ratsam, den sachkundigen Benutzer bzw. die Benutzerin stark einzubinden, das heißt insbesondere, das vorhandene Wissen der Wissenschaftler ausnutzen. Softwarelösungen können zwar Ähnlichkeitssuchen zu bekannten Krankheitsgenen durchführen oder Gene nach anderen Parametern sortieren, sie sind jedoch derzeit noch nicht in der Lage, ebenso wie Menschen Assoziationen zu folgen und flexibel aufgrund von Zwischenergebnissen die Prioritisitierungsstrategie zu ändern. Anforderung an eine Suchmaschine war es somit, die Benutzer von unnötiger Arbeit zu entlasten und gleichzeitig eine permanente interaktive Veränderung der Suchbedingungen ermöglichen.
GeneDistiller wurde deshalb so programmiert, dass den Wissenschaftlern umfangreiche genspezifische Daten aus diversen Quellen zur Verfügung stehen. Um aus dieser Fülle von Daten die gewünschten Informationen zu erhalten, besteht die Möglichkeit der Sortierung, Selektion (Filterung nach Eigenschaften der Gene u.ä.) und der Projektion (Auswahl der anzuzeigenden Datenquellen). Darüber hinaus kann auch weniger stringent vorgegangen werden und lediglich eine optische Hervorhebung bestimmter Eigenschaften gewählt werden. *GeneDistiller* erlaubt es fernerhin, bekannte Krankheiten auszuwählen und nach Genen zu suchen, die mit bekannten Krankheitsgenen interagieren oder ähnliche Eigenschaften wie diese aufweisen. Neben der Selektion bzw. optischen Hervorhebung bietet *GeneDistiller* auch die Möglichkeit einer Prioritisierung der Gene nach den Wünschen der Benutzer. Hierbei handelt es sich nicht um eine einfache Sortierung, sondern es werden Punkte für die Erfüllung aller angegebenen Kriterien (im Falle einer Ähnlichkeitssuche z.B. Interaktion mit bekannten Genen, ähnliche GeneOntology Einträge (17), ähnliche HPO-Einträge (5) usw. sowie weitere benutzerdefinierte Angaben wie z.B. das Vorhandensein bestimmter Suchbegriffe in den OMIM-Angaben (18) zu den Genen) vergeben. Die Prioritisierungsstrategie (z.B. Suche nach mitochondrialen Proteinen oder nach in einem oder mehreren Geweben exprimierten Genen) bestimmt dabei die Wichtung der einzelnen Punkte. Für häufige Ansätze existieren fertige Profile, die aber jederzeit von den Benutzern verändert werden können. Da eine typische Abfrage in wenigen Sekunden präsentiert wird, bietet sich so die Möglichkeit, die Strategie aufgrund der Ergebnisse sofort zu verändern, und so interaktiv Hypothesen zu verfeinern oder neue aufzustellen und zu überprüfen. Zu allen dargestellten Daten werden Hyperlinks zu den Originaldaten angeboten, so dass ein ‘drill-down’ ermöglicht wird.
GeneDistiller lässt sich jedoch nicht nur mit genetischen Regionen aufrufen, sondern es können auch gezielt Kandidatengenlisten übermittelt werden. Für diese stehen dann sämtliche Funktionen von *GeneDistiller* ebenso zur Verfügung wie für positionelle Kandidaten.
HomozygosityMapper und *GeneDistiller* sind eng verzahnt; zu den mit *HomozygosityMapper* bestimmten Regionen werden direkte Links zu *GeneDistiller* angeboten. Neben den schon genannten Eigenschaften der Gene kann so auch die Homozygotie in einer Studie als Kriterium herangezogen werden. Ebenso lassen sich (genomweite) Kandidatengenlisten gezielt auf Homozygotie überprüfen und entsprechend sortieren und filtern.

Implementierung

Alle Applikationen basieren auf frei nutzbaren *open source* Programmen. Als Datenbanksystem wurde *PostgreSQL 8* benutzt, sämtliche Programme wurden in *Perl 5.8* programmiert und alle Web-Applikationen laufen unter einem *Apache 2.2* Webserver auf Servern mit *Debian 5* bzw. *Fedora Core 10* Linux. Für die Web Interfaces werden klassische HTML Seiten benutzt, die zum Teil durch JavaScript und Ajax benutzerfreundlicher gestaltet wurden. Sie wurden mit *Mozilla Firefox 3* entwickelt und optimiert, funktionieren aber auch mit allen anderen modernen Internetbrowsern.

Eine generelle Anforderung bei der Erstellung der Programme war die unkomplizierte Nutzbarkeit durch die Anwenderinnen und Anwender. Web-basierte Programme bieten hier generell den Vorteil, dass keinerlei lokale Installationen erforderlich sind und dass die Benutzung eines Internetbrowsers und von Internet-Formularen allgemein bekannt ist. Nachteilig ist dabei jedoch, dass sämtliche Rechenleistung von einem Charité-Server erbracht werden muss. Außerdem muss sichergestellt sein, dass Wissenschaftler ihre Daten vor unberechtigtem Zugriff durch Dritte schützen können. Es sind somit Authentifizierungsüberprüfungen erforderlich. In Abwägung dieser Problematik wurde *AssociationDB* (19) als eine von den Anwendern lokal zu installierende Datenbank realisiert (die allerdings die Installation eines Webservers, eines Datenbankmanagementsystems sowie von Perl voraussetzt) und *FragIdent* (20) als Datei-basierte Anwendung, die unter Windows ohne jegliche weitere Installation benutzt werden kann.

Alle anderen Programme sind web-basiert und laufen auf Servern der Charité. Da im Falle von *HomozygosityMapper* ebenfalls gewährleistet sein musste, dass lediglich die Besitzer der Daten diese einsehen können, gleichzeitig aber Kooperationspartner einige dieser Datensätze nutzen können sollten, wurde hier ein Rechtesystem implementiert, das den Besitzern der Daten gestattet, den Zugriff auf diese granulär zu bestimmen. Nach Veröffentlichung können die Projekte so auch öffentlich gemacht werden; *HomozygosityMapper* fungiert somit als ein öffentliches *repository* für Genotypen – diese öffentliche Verfügbarkeit der Originaldaten wird beispielsweise im Bereich Genexpression bereits von vielen Zeitschriften als Bedingung für eine Veröffentlichung gefordert.

Da eine Gensuche nach komplexen Vorgaben umfangreiche Eingaben in *GeneDistiller* erfordern kann, wurde hier eine Lösung über speicherbare Hyperlinks gewählt. Mit diesen können Benutzer jederzeit das Interface mitsamt aller Eingaben wiederherstellen und die Hyperlinks zudem auch an Kooperationspartner übermitteln, die unabhängig von ihnen mit diesen weiterarbeiten können.

Sämtliche Applikation wurden iterativ entwickelt. Das heißt, dass spätere Nutzer und Nutzerinnen schon während der Entwicklung mit den Programmen arbeiteten und die Programme anhand ihrer Vorschläge modifiziert werden konnten.

Ergebnisse

Im Rahmen meiner Promotion habe ich insgesamt 5 Applikationen entwickelt bzw. maßgeblich daran mitgearbeitet:

- *HomozygosityMapper* (2) (http://www.homozygositymapper.org)
- *GeneDistiller* (4) (http://www.genedistiller.org)
- *AssociationDB* (19) (http://compbio.charite.de/genetik/AssociationDB/)
- *FragIdent* (20) (http://compbio.charite.de/genetik/FragIdent/)
- *The Human Phenotype Ontology* (5) (http://www.human-phenotype-ontology.org)

Außerdem ist eine Studie zum Einsatz der Homozygotiekartierung in Familien ohne offensichtlichen konsanguinen Hintergrund (21) entstanden.
Frühe Entwicklungsstufen von *HomozygosityMapper* wurden parallel zum konventionellen Ansatz in einigen Homozygotiekartierungen eingesetzt (22-27).

Diskussion

In den untersuchten Homozygotiekartierungen lieferte *HomozygosityMapper* stets dieselben Regionen wie eine konventionelle Kopplungsanalyse; allerdings um Größenordnungen schneller. In einem Fall konnte die Geschwindigkeit um das 24.000-fache gesteigert werden (ca. 5 Minuten statt 2.000 Stunden bei einer großen konsanguinen Familie und allen verfügbaren 50.000 SNPs). Der konventionelle Ansatz ist *HomozygosityMapper* hingegen dann überlegen, wenn die Information nicht betroffener Familienmitgliedern genutzt werden kann. Allerdings ist auch hier durch die Kombination beider Verfahren ein beträchtlicher Geschwindigkeitszuwachs möglich, wenn zunächst homozygote Regionen durch *HomozygosityMapper* identifiziert werden und anstelle des gesamten Genoms lediglich diese homozygoten Regionen mittels Kopplungsanalyse untersucht werden. Überlegungen, homozygote Regionen in gesunden Geschwistern automatisch auszuschließen, wurden verworfen, da kurze homozygote Segmente in diesen einerseits durch die geringe Informativität von SNPs zufällig entstehen können und andererseits im Falle genetischer Heterogenität so für einige Familien relevante Regionen fälschlich ausgeschlossen würden (Typ II Fehler, falsch negativ). Stattdessen wurde in *HomozygosityMapper* inzwischen eine Funktion eingefügt, mit der eine Zweipunkt-Kopplungsanalyse für eine homozygote Region durchgeführt werden kann und Dateien für eine Mehrpunkt-Analyse mit Allegro (8) exportiert werden können. Darüber hinaus kann eine Region unter Umständen auch durch die einfache visuelle Inspektion der Genotypen in *HomozygosityMapper* ausgeschlossen werden.

Literaturverzeichnis

Die in diese Arbeit eingehenden 6 Publikationen werden auf Seite 14 noch einmal explizit und unter Angabe der „5-year impact factors“ (ISI Web of Knowledge) aufgelistet und sind ab Seite 15 in dieser Arbeit enthalten (in der elektronischen Version nur als Abstracts).

1. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308-311.

2. Seelow D., Schuelke,M., Hildebrandt,F. and Nurnberg,P. (2009) HomozygosityMapper - an interactive approach to homozygosity mapping. *Nucleic Acids Res.*, **doi:10.1093/nar/gkp369**.

3. The International HapMap Project (2003) The International HapMap Project. *Nature*, **426**, 789-796.

4. Seelow,D., Schwarz,J.M. and Schuelke,M. (2008) GeneDistiller--distilling candidate genes from linkage intervals. *PLoS. ONE.*, **3**, e3874.

5. Robinson,P.N., Kohler,S., Bauer,S., Seelow,D., Horn,D. and Mundlos,S. (2008) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.*, **83**, 610-615.

6. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365-370.

7. Sobel,E., Sengul,H. and Weeks,D.E. (2001) Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees. *Hum. Hered.*, **52**, 121-131.

8. Gudbjartsson,D.F., Thorvaldsson,T., Kong,A., Gunnarsson,G. and Ingolfsdottir,A. (2005) Allegro version 2. *Nat. Genet.*, **37**, 1015-1016.

9. Kruglyak,L., Daly,M.J., Reeve-Daly,M.P. and Lander,E.S. (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.*, **58**, 1347-1363.

10. Kong,A. and Cox,N.J. (1997) Allele-sharing models: LOD scores and accurate linkage tests. *Am. J. Hum. Genet.*, **61**, 1179-1188.

11. Strauch,K., Furst,R., Ruschendorf,F., Windemuth,C., Dietter,J., Flaquer,A., Baur,M.P. and Wienker,T.F. (2005) Linkage analysis of alcohol dependence using MOD scores. *BMC. Genet.*, **6 Suppl 1**, S162.

12. Lander,E.S. and Botstein,D. (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science*, **236**, 1567-1570.

13. Goedken,R., Ludington,E., Crowe,R., Fyer,A.J., Hodge,S.E., Knowles,J.A., Vieland,V.J. and Weissman,M.M. (2000) Drawbacks of GENEHUNTER for larger pedigrees: application to panic disorder. *Am. J. Med. Genet.*, **96**, 781-783.

14. Kirk,K.M. and Cardon,L.R. (2002) The impact of genotyping error on haplotype reconstruction and frequency estimation. *Eur. J. Hum. Genet.*, **10**, 616-622.

15. Woods,C.G., Valente,E.M., Bond,J. and Roberts,E. (2004) A new method for autozygosity mapping using single nucleotide polymorphisms (SNPs) and EXCLUDEAR. *J. Med. Genet.*, **41**, e101.

16. Woods,C.G., Cox,J., Springell,K., Hampshire,D.J., Mohamed,M.D., McKibbin,M., Stern,R., Raymond,F.L., Sandford,R., Malik,S.S. *et al.* (2006) Quantification of homozygosity in consanguineous individuals with autosomal recessive disease. *Am. J. Hum. Genet.*, **78**, 889-896.

17. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258-D261.

18. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514-D517.

19. Seelow,D., Hoffmann,K. and Lindner,T.H. (2007) AssociationDB: web-based exploration of genomic association. *Bioinformatics.*, **23**, 2643-2644.

20. Seelow,D., Goehler,H. and Hoffmann,K. (2009) FragIdent - Automatic identification and characterisation of cDNA-fragments. *BMC. Genomics*, **10**, 95.

21. Hildebrandt,F., Heeringa,S.F., Ruschendorf,F., Attanasio,M., Nurnberg,G., Becker,C., Seelow,D., Huebner,N., Chernin,G., Vlangos,C.N. *et al.* (2009) A systematic approach to mapping recessive disease genes in individuals from outbred populations. *PLoS. Genet.*, **5**, e1000353.

22. Attanasio,M., Uhlenhaut,N.H., Sousa,V.H., O'Toole,J.F., Otto,E., Anlag,K., Klugmann,C., Treier,A.C., Helou,J., Sayer,J.A. *et al.* (2007) Loss of GLIS2 causes nephronophthisis in humans and mice by increased apoptosis and fibrosis. *Nat. Genet.*, **39**, 1018-1024.

23. Bergmann,C., Senderek,J., Anhuf,D., Thiel,C.T., Ekici,A.B., Poblete-Gutierrez,P., van,S.M., Seelow,D., Nurnberg,G., Schild,H.H. *et al.* (2006) Mutations in the gene encoding the Wnt-signaling component R-spondin 4 (RSPO4) cause autosomal recessive anonychia. *Am. J. Hum. Genet.*, **79**, 1105-1109.

24. Hinkes,B., Wiggins,R.C., Gbadegesin,R., Vlangos,C.N., Seelow,D., Nurnberg,G., Garg,P., Verma,R., Chaib,H., Hoskins,B.E. *et al.* (2006) Positional cloning uncovers mutations in PLCE1 responsible for a nephrotic syndrome variant that may be reversible. *Nat. Genet.*, **38**, 1397-1405.

25. Jenkins,D., Seelow,D., Jehee,F.S., Perlyn,C.A., Alonso,L.G., Bueno,D.F., Donnai,D., Josifova,D., Mathijssen,I.M., Morton,J.E. *et al.* (2007) RAB23 mutations in Carpenter syndrome imply an unexpected role for hedgehog signaling in cranial-suture development and obesity. *Am. J. Hum. Genet.*, **80**, 1162-1170.

26. Konrad,M., Schaller,A., Seelow,D., Pandey,A.V., Waldegger,S., Lesslauer,A., Vitzthum,H., Suzuki,Y., Luk,J.M., Becker,C. *et al.* (2006) Mutations in the tight-junction gene claudin 19 (CLDN19) are associated with renal magnesium wasting, renal failure, and severe ocular involvement. *Am. J. Hum. Genet.*, **79**, 949-957.

27. Cizkova,A., Stranecky,V., Mayr,J.A., Tesarova,M., Havlickova,V., Paul,J., Ivanek,R., Kuss,A.W., Hansikova,H., Kaplanova,V. *et al.* (2008) TMEM70 mutations cause isolated ATP synthase deficiency and neonatal mitochondrial encephalocardiomyopathy. *Nat. Genet.*, **40**, 1288-1290.

Ausgewählte Publikationen / Anteilserklärung

1. **Seelow,D**., Hoffmann,K. and Lindner,T.H. (2007) AssociationDB: web-based exploration of genomic association. Bioinformatics., 23, 2643-2644.
   (5-year impact factor: 6.649)
   Die Software basiert auf einer Idee von D. Seelow und wurde komplett von ihm entwickelt (Anteil 80 Prozent).

2. **Seelow,D**., Schwarz,J.M. and Schuelke,M. (2008) GeneDistiller - distilling candidate genes from linkage intervals. PLoS. ONE., 3, e3874.
   (bislang kein Impact Factor)
   D. Seelow integrierte die externen Daten und entwickelte die Software (Anteil 80 Prozent).

3. Robinson,P.N., Kohler,S., Bauer,S., **Seelow,D**., Horn,D. and Mundlos,S. (2008) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am. J. Hum. Genet., 83, 610-615.
   (5-year impact factor: 11.711)
   D. Seelow erstellte die zugrunde liegende Datenbank sowie web-basierte Benutzerschnittstellen (Anteil 5 Prozent).

4. **Seelow,D**., Goehler,H. and Hoffmann,K. (2009) FragIdent - Automatic identification and characterisation of cDNA-fragments. BMC. Genomics, 10, 95.
   (5-year impact factor: 4.243)
   Die Software basiert auf einer Idee von D. Seelow und wurde komplett von ihm entwickelt (Anteil 80 Prozent).

5. Hildebrandt,F., Heeringa,S.F., Ruschendorf,F., Attanasio,M., Nurnberg,G., Becker,C., **Seelow,D**., Huebner,N., Chernin,G., Vlangos,C.N. et al. (2009) A systematic approach to mapping recessive disease genes in individuals from outbred populations. PLoS. Genet., 5, e1000353.
   (5-year impact factor: 8.733)
   In diese Arbeit flossen Datenanalysen von D. Seelow sowie Vorschläge zur Heransgehensweise ein (Anteil 8 Prozent).

6. **Seelow D**., Schuelke,M., Hildebrandt,F. and Nurnberg,P. (2009) HomozygosityMapper - an interactive approach to homozygosity mapping. Nucleic Acids Res., doi:10.1093/nar/gkp369.
   (5-year impact factor: 7.163)
   Die Software basiert auf einer Idee von D. Seelow und wurde komplett von ihm entwickelt (Anteil 80 Prozent).

Dominik Seelow
Promovend

Prof. Dr. Markus Schülke
betreuender Hochschullehrer

AssociationDB: web-based exploration of genomic association.

Seelow D, Hoffmann K, Lindner TH.

Genome-wide association studies use hundreds of thousands of markers making it challenging to present and finally interpret the results. We developed a graphical, web-based solution for an interactive exploration of the results of case-control studies, with a tight integration of related gene information and tissue-specific expression data. Association results are presented as physical position-based vertical bars with known genes included as horizontal bars at their respective physical positions. The interface allows the specification of filtering criteria for the association data and highlights potentially interesting genes with user-specified terms occurring in their reports or with relevant expression patterns. Pop-up windows and hyperlinks provide drill-down capabilities and quick access to relevant data AssociationDB can either be used as a stand-alone solution or as a front-end joining association results obtained by other software with genomic information.

Bioinformatics. 2007 Oct 1;23(19):2643-4. Epub 2007 Jul 27.

GeneDistiller - distilling candidate genes from linkage intervals.

Seelow D, Schwarz JM, Schuelke M.

BACKGROUND: Linkage studies often yield intervals containing several hundred positional candidate genes. Different manual or automatic approaches exist for the determination of the gene most likely to cause the disease. While the manual search is very flexible and takes advantage of the researchers' background knowledge and intuition, it may be very cumbersome to collect and study the relevant data. Automatic solutions on the other hand usually focus on certain models, remain "black boxes" and do not offer the same degree of flexibility.
METHODOLOGY: We have developed a web-based application that combines the advantages of both approaches. Information from various data sources such as gene-phenotype associations, gene expression patterns and protein-protein interactions was integrated into a central database. Researchers can select which information for the genes within a candidate interval or for single genes shall be displayed. Genes can also interactively be filtered, sorted and prioritised according to criteria derived from the background knowledge and preconception of the disease under scrutiny.
CONCLUSIONS: GeneDistiller provides knowledge-driven, fully interactive and intuitive access to multiple data sources. It displays maximum relevant information, while saving the user from drowning in the flood of data. A typical query takes less than two seconds, thus allowing an interactive and explorative approach to the hunt for the candidate gene.
ACCESS: GeneDistiller can be freely accessed at http://www.genedistiller.org/.

PLoS One. 2008;3(12):e3874. Epub 2008 Dec 5.

The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease.

Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S.

There are many thousands of hereditary diseases in humans, each of which has a specific combination of phenotypic features, but computational analysis of phenotypic data has been hampered by lack of adequate computational data structures. Therefore, we have developed a Human Phenotype Ontology (HPO) with over 8000 terms representing individual phenotypic anomalies and have annotated all clinical entries in Online Mendelian Inheritance in Man with the terms of the HPO. We show that the HPO is able to capture phenotypic similarities between diseases in a useful and highly significant fashion.

Am J Hum Genet. 2008 Nov;83(5):610-5. Epub 2008 Oct 23.

FragIdent - automatic identification and characterisation of cDNA-fragments.

Seelow D, Goehler H, Hoffmann K.

BACKGROUND: Many genetic studies and functional assays are based on cDNA fragments. After the generation of cDNA fragments from an mRNA sample, their content is at first unknown and must be assigned by sequencing reactions or hybridisation experiments. Even in characterised libraries, a considerable number of clones are wrongly annotated. Furthermore, mix-ups can happen in the laboratory. It is therefore essential to the relevance of experimental results to confirm or determine the identity of the employed cDNA fragments. However, the manual approach for the characterisation of these fragments using BLAST web interfaces is not suited for larger number of sequences and so far, no user-friendly software is publicly available.
RESULTS: Here we present the development of FragIdent, an application for the automatic identification of open reading frames (ORFs) within cDNA-fragments. The software performs BLAST analyses to identify the genes represented by the sequences and suggests primers to complete the sequencing of the whole insert. Gene-specific information as well as the protein domains encoded by the cDNA fragment are retrieved from Internet-based databases and included in the output. The application features an intuitive graphical interface and is designed for researchers without any bioinformatics skills. It is suited for projects comprising up to several hundred different clones.
CONCLUSION: We used FragIdent to identify 84 cDNA clones from a yeast two-hybrid experiment. Furthermore, we identified 131 protein domains within our analysed clones. The source code is freely available from our homepage at
http://compbio.charite.de/genetik/FragIdent/.

BMC Genomics. 2009 Mar 2;10:95.

A systematic approach to mapping recessive disease genes in individuals from outbred populations.

Hildebrandt F, Heeringa SF, Rüschendorf F, Attanasio M, Nürnberg G, Becker C, Seelow D, Huebner N, Chernin G, Vlangos CN, Zhou W, O'Toole JF, Hoskins BE, Wolf MT, Hinkes BG, Chaib H, Ashraf S, Schoeb DS, Ovunc B, Allen SJ, Vega-Warner V, Wise E, Harville HM, Lyons RH, Washburn J, Macdonald J, Nürnberg P, Otto EA.

The identification of recessive disease-causing genes by homozygosity mapping is often restricted by lack of suitable consanguineous families. To overcome these limitations, we apply homozygosity mapping to single affected individuals from outbred populations. In 72 individuals of 54 kindred ascertained worldwide with known homozygous mutations in 13 different recessive disease genes, we performed total genome homozygosity mapping using 250,000 SNP arrays. Likelihood ratio Z-scores (ZLR) were plotted across the genome to detect ZLR peaks that reflect segments of homozygosity by descent, which may harbor the mutated gene. In 93% of cases, the causative gene was positioned within a consistent ZLR peak of homozygosity. The number of peaks reflected the degree of inbreeding. We demonstrate that disease-causing homozygous mutations can be detected in single cases from outbred populations within a single ZLR peak of homozygosity as short as 2 Mb, containing an average of only 16 candidate genes. As many specialty clinics have access to cohorts of individuals from outbred populations, and as our approach will result in smaller genetic candidate regions, the new strategy of homozygosity mapping in single outbred individuals will strongly accelerate the discovery of novel recessive disease genes.

PLoS Genet. 2009 Jan;5(1):e1000353. Epub 2009 Jan 23.

HomozygosityMapper - an interactive approach to homozygosity mapping.

Seelow D, Schuelke M, Hildebrandt F, Nürnberg P.

Homozygosity mapping is a common method for mapping recessive traits in consanguineous families. In most studies, applications for multipoint linkage analyses are applied to determine the genomic region linked to the disease. Unfortunately, these are neither suited for very large families nor for the inclusion of tens of thousands of SNPs. Even if less than 10,000 markers are employed, such an analysis may easily last hours if not days. Here we present a web-based approach to homozygosity mapping. Our application stores marker data in a database into which users can directly upload their own SNP genotype files. Within a few minutes, the database analyses the data, detects homozygous stretches and provides an intuitive graphical interface to the results. The homozygosity in affected individuals is visualized genome-wide with the ability to zoom into single chromosomes and user-defined chromosomal regions. The software also displays the underlying genotypes in all samples. It is integrated with our candidate gene search engine, GeneDistiller, so that users can interactively determine the most promising gene. They can at any point restrict access to their data or make it public, allowing HomozygosityMapper to be used as a data repository for homozygosity-mapping studies. HomozygosityMapper is available at http://www.homozygositymapper.org/.

Nucleic Acids Res. 2009 Jul 1;37(Web Server issue):W593-9. Epub 2009 May 21.

Lebenslauf

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

Publikationsliste

1. Stober,G., **Seelow,D.**, Ruschendorf,F., Ekici,A., Beckmann,H. and Reis,A. (2002) Periodic catatonia: confirmation of linkage to chromosome 15 and further evidence for genetic heterogeneity. *Hum. Genet.*, **111**, 323-330.

2. Kaynak,B., von,H.A., Mebus,S., **Seelow,D.**, Hennig,S., Vogel,J., Sperling,H.P., Pregla,R., exi-Meskishvili,V., Hetzer,R. *et al.* (2003) Genome-wide array analysis of normal and malformed human hearts. *Circulation*, **107**, 2467-2474.

3. **Seelow,D.**, Galli,R., Mebus,S., Sperling,H.P., Lehrach,H. and Sperling,S. (2004) d-matrix - database exploration, visualization and analysis. *BMC. Bioinformatics.*, **5**, 168.

4. Bergmann,C., Senderek,J., Anhuf,D., Thiel,C.T., Ekici,A.B., Poblete-Gutierrez,P., van,S.M., **Seelow,D.**, Nurnberg,G., Schild,H.H. *et al.* (2006) Mutations in the gene encoding the Wnt-signaling component R-spondin 4 (RSPO4) cause autosomal recessive anonychia. *Am. J. Hum. Genet.*, **79**, 1105-1109.

5. Hinkes,B., Wiggins,R.C., Gbadegesin,R., Vlangos,C.N., **Seelow,D.**, Nurnberg,G., Garg,P., Verma,R., Chaib,H., Hoskins,B.E. *et al.* (2006) Positional cloning uncovers mutations in PLCE1 responsible for a nephrotic syndrome variant that may be reversible. *Nat. Genet.*, **38**, 1397-1405.

6. Konrad,M., Schaller,A., **Seelow,D.**, Pandey,A.V., Waldegger,S., Lesslauer,A., Vitzthum,H., Suzuki,Y., Luk,J.M., Becker,C. *et al.* (2006) Mutations in the tight-junction gene claudin 19 (CLDN19) are associated with renal magnesium wasting, renal failure, and severe ocular involvement. *Am. J. Hum. Genet.*, **79**, 949-957.

7. **Seelow,D.**, Hoffmann,K. and Lindner,T.H. (2007) AssociationDB: web-based exploration of genomic association. *Bioinformatics.*, **23**, 2643-2644.

8. Attanasio,M., Uhlenhaut,N.H., Sousa,V.H., O'Toole,J.F., Otto,E., Anlag,K., Klugmann,C., Treier,A.C., Helou,J., Sayer,J.A., **Seelow,D.**, Nürnberg,G., Becker,C., Chudley,A.E., Nürnberg,P., Hildebrandt,F., Treier,M. (2007) Loss of GLIS2 causes nephronophthisis in humans and mice by increased apoptosis and fibrosis. *Nat. Genet.*, **39**, 1018-1024.

9. Jenkins,D., **Seelow,D.**, Jehee,F.S., Perlyn,C.A., Alonso,L.G., Bueno,D.F., Donnai,D., Josifova,D., Mathijssen,I.M., Morton,J.E. *et al.* (2007) RAB23 mutations in Carpenter syndrome imply an unexpected role for hedgehog signaling in cranial-suture development and obesity. *Am. J. Hum. Genet.*, **80**, 1162-1170.

10. **Seelow,D.**, Schwarz,J.M. and Schuelke,M. (2008) GeneDistiller--distilling candidate genes from linkage intervals. *PLoS. ONE.*, **3**, e3874.

11. Robinson,P.N., Kohler,S., Bauer,S., **Seelow,D.**, Horn,D. and Mundlos,S. (2008) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.*, **83**, 610-615.

12. Michalk,A., Stricker,S., Becker,J., Rupps,R., Pantzar,T., Miertus,J., Botta,G., Naretto,V.G., Janetzki,C., Yaqoob,N. Ott,C.E., **Seelow,D.**, Wieczorek,D., Fiebig,B., Wirth,B., Hoopmann,M., Walther,M., Körber,F., Blankenburg,M., Mundlos,S., Heller,R., Hoffmann,K. (2008) Acetylcholine receptor pathway mutations explain various fetal akinesia deformation sequence disorders. *Am. J. Hum. Genet.*, **82**, 464-476.

13. **Seelow,D.**, Schuelke,M., Hildebrandt,F. and Nurnberg,P. (2009) HomozygosityMapper--an interactive approach to homozygosity mapping. *Nucleic Acids Res.*, **doi:10.1093/nar/gkp369**.

14. **Seelow,D.**, Goehler,H. and Hoffmann,K. (2009) FragIdent--automatic identification and characterisation of cDNA-fragments. *BMC. Genomics*, **10**, 95.

15. Hildebrandt,F., Heeringa,S.F., Ruschendorf,F., Attanasio,M., Nurnberg,G., Becker,C., **Seelow,D.**, Huebner,N., Chernin,G., Vlangos,C.N. *et al.* (2009) A systematic approach to mapping recessive disease genes in individuals from outbred populations. *PLoS. Genet.*, **5**, e1000353.

Selbständigkeitserklärung

„Ich, Dominik Seelow, erkläre, dass ich die vorgelegte Dissertation mit dem Thema: **Erstellung eines computergestützten Verfahrens zur Suche nach Kandidatengenen für rezessiv vererbte Krankheiten in konsanguinen Familien** selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, ohne die (unzulässige) Hilfe Dritter verfasst und auch in Teilen keine Kopien anderer Arbeiten dargestellt habe.“

Datum Unterschrift

Danksagung

Am Zustandekommen dieser Dissertation bzw. der ihr zugrunde liegenden Publikationen waren zahlreiche Kolleginnen und Kollegen beteiligt. Unter ihnen möchte ich die folgenden Personen besonders hervorheben:

Jana Marie Schwarz,
der es immer wieder gelungen ist, mich zu motivieren, auch langweilige Aufgaben zu erledigen und meine manchmal etwas wirr formulierten Ideen irgendwie verständlich darzustellen,

Katrin Hoffmann,
die zahlreiche neue Ideen einbrachte und mich immer wieder ermunterte, Software nicht nur zu entwickeln sondern auch in wissenschaftlichen Publikationen zu veröffentlichen,

Gudrun Nürnberg,
die nahezu sämtliche Versionen von HomozygosityMapper ausgiebig getestet, benutzt und kritisiert hat und ohne deren Verbesserungsvorschläge die Software nie so weit gekommen wäre,

Peter Nürnberg,
der immer an mich geglaubt und mich in allen Dingen unterstützt hat,

und Markus Schülke
für die vielen guten Ideen, sein Vertrauen in mich und vor allem für das äußert angenehme Arbeitsklima.