GWAS am Menschen

Historische Perspektive, Motivation und „*common disease-common variant*“-Hypothese

GWAS steht für genomweite Assoziationsstudie bzw. Scan und beschreibt eine relativ neue, aber sehr einflussreiche Technik zur Untersuchung der genetischen Faktoren, die menschliche Phänotypen beeinflussen.

Dass viele menschliche Eigenschaften, wie z.B. Körpergrösse oder Augenfarbe, vererbt werden, wissen wir aus unserer persönlichen Erfahrung. Aber auch die Anfälligkeit für viele Krankheiten (Krebs, Herzkreislauferkrankungen, Typ-2-Diabetes etc.) ist stark erblich. Für einige dieser Krankheiten, insbesondere Krankheiten, die von einem einzigen Gen beeinflusst werden (z.B. Sichelzellanämie oder Blutgerinnungserkrankungen), war es auch schon länger möglich, die genetischen Grundlagen zu bestimmen.

Aber gerade für häufig auftretende, quantitative Phänotypen (z.B. Bluthochdruck, Krebsrisiko) war es beim Menschen sehr schwer, die zugrundeliegenden genetischen Faktoren umfassend zu bestimmen. Der Grund dafür ist, dass solche quantitativen Phänotypen in der Regel durch mehrere Gene beeinflusst werden und der Einfluss jedes einzelnen Gens relativ klein ist. Dadurch sind klassische Mapping-Ansätze nur schwer anwendbar und der Ansatz von Sättigungsmutageneseexperimenten, wie sie bei Modellorganismen angewendet wird, verbietet sich aus offensichtlichen ethischen Gründen.

Kandidaten/Gen-Studien

Mit der Einführung automatisierter DNA-Sequenziertechniken in den 90er Jahren wurden die sogenannten Kandidaten-Gen-Assoziationsstudien beliebt. In diesen Studien wählt man basierend auf mechanistischen Überlegungen ein Gen aus, von dem man vermutet, dass dieses den Phänotyp beeinflussen könnte (z.B. das Insulin Rezeptorgen *INSR* für eine Studie an Diabetes) und prüft dann, ob genetische Variationen in diesem Gen mit dem Phänotyp assoziiert sind.

Offensichtlich ist dieser Ansatz nicht geeignet um neue Gene zu identifizieren, von denen man nicht ohnehin schon annimmt, dass sie den Phänotyp beeinflussen.

Mit der Sequenzierung des menschlichen Genoms wurde der gesamte Katalog der menschlichen Gene bekannt und es wurde daher möglich, ohne vorherige Annahme (Kandidaten-Gen-Ansatz) die Analyse auf alle menschlichen Gene bzw. das ganze Genom anzuwenden.

Seltene Krankheiten vs. common disease-common variant

Mit der systematischen Sequenzierung menschlicher Gene wurde schnell offensichtlich, dass ein Grossteil der interindividuellen Variation in Form von SNPs, die in der Gesamtbevölkerung weit verbreitet sind, auftritt. Diese weitverbreiteten genetischen Variationen (engl. *common variants*) sind das Gegenteil von seltenen oder privaten Variationen (engl. *rare* oder *private variants*), die nur in einer ganz kleinen Gruppe oder in einer bestimmten Einzelperson auftreten.

Seltene Krankheiten (d.h. Krankheiten, die weniger als 1 in 2000 Menschen betrifft) haben ihre Ursache oft in einer einzelnen, für eine bestimmte Familie spezifischen genetischen Variation. Diese Tatsache vereinfacht das Mapping dieser Mutationen mit klassischen genetischen Methoden. Die Tay-Sachs-Krankheit ist ein klassisches Beispiel einer seltenen Krankheit. In dieser Krankheit inaktivieren Mutationen das Gen für das Enzym beta-Hexosaminidase A. Dadurch kommt es zu einer toxischen Anrei-

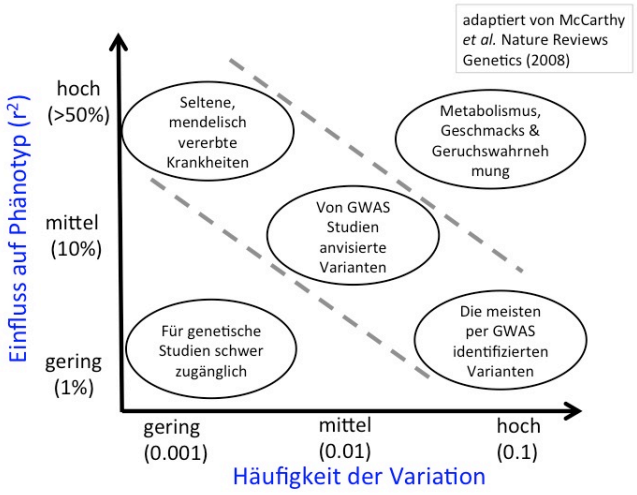

Abbildung 1 *Seltene Variationen mit starkem Einfluss können gut mithilfe von klassischen Familienstudien untersucht werden. Häufiger auftretende Varianten mit geringerem Effekt können besser durch GWAS-Studien untersucht werden. Variationen, die sowohl häufig sind, als auch einen grossen Einfluss auf den Phänotyp haben, sind im medizinischen Bereich sehr selten, treten in anderen Bereichen (z.B. Metabolismus, Sinneswahrnehmung etc.) aber durchaus auf. Seltene Varianten mit geringem Einfluss sind für genetische Studien nur schwer zugänglich, aber ohnehin eher uninteressant.*

cherung des Substrats und zur Schädigung von Nervenzellen. Wie für seltene Krankheiten üblich, haben unterschiedliche Tay-Sachs-Patienten in der Regel unterschiedliche Mutationen (fast 100 unterschiedliche Mutationen sind bekannt). Alle diese Mutationen betreffen aber das gleiche Enzym. Wie schwerwiegend die Krankheit ist, hängt dabei davon ab, wie stark die Mutation die Funktion des Enzyms einschränkt.

Im Gegensatz dazu steht die „*common disease-common variant*“-Hypothese für weitverbreitete Krankheiten. Diese Hypothese besagt, dass weitverbreitete Krankheiten (z.B. Herz-Kreislauf-Erkrankungen, Krebs, Diabetes) auch von weitverbreiteten genetischen Variationen beeinflusst werden. Ursache für ein besonders hohes Risiko wäre dabei nicht eine besonders schwerwiegende Mutation in einem einzigen „Risiko-Gen“, sondern das Zusammentreffen mehrerer ungünstiger Variationen in verschiedenen Genen. Dabei kann jede einzelne dieser Variationen recht häufig auftreten. GWAS-Studien wurden entwickelt, um solche „*common variant-common disease*“-Zusammenhänge zu untersuchen, die durch herkömmliche genetische Studien nur schwer erfasst werden können.

Die Grundidee von GWAS-Studien

GWAS-Studien untersuchen statistische Zusammenhänge zwischen genotypischer und phänotypischer Variation mit dem Ziel, solche genetischen Variationen zu finden, die einen bestimmten biologischen Prozess beeinflussen.

In einer GWAS-Studie macht man *a priori* keine Annahme darüber, welche Gene an dem untersuchten Prozess beteiligt sind. Der eigentliche Startpunkt einer GWAS Studie ist also der Phänotyp. In diesem Sinne ist der Informationsfluss in einer GWAS-Studie (vom Phänotyp zum Genotyp) exakt umgekehrt zum biologischen Informationsfluss, wo die Information ja vom Genotyp zum Phänotyp fliesst.

Gendaten für GWAS

Neben dem Phänotyp braucht man Informationen über die genetischen Variationen der Studienteilnehmer. Diese Information misst man im Regelfall mit sogenannten Genotypisierungs-Microarrays, die in einem einzigen Experiment den Genotyp einer Person an ca. 1 Million SNPs misst. Diese SNPs sind dabei von dem Chip-Hersteller so ausgewählt, dass in jedem LD-Block des Genoms einige SNPs gemessen werden. Aufgrund der Struktur der genetischen Variation beim Menschen reicht diese Information aus, um den Genotyp an der überwiegenden Mehrzahl aller menschlichen SNPs abzuleiten.

Die Genotypinformation nimmt dann letztendlich die Form einer langen Liste von SNPs mit dem jeweils dazugehörigen Genotyp an.

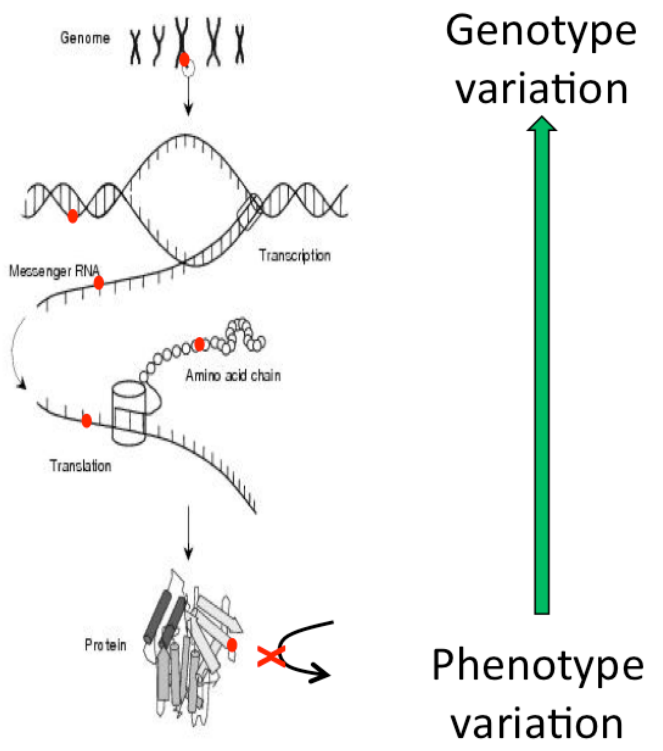

Abbildung 2 *In GWAS-Studien sucht man ausgehend vom Phänotyp nach den genetischen Variationen, die diesen Phänotyp beeinflussen.*

Die Kosten für eine solche Genotypisierung sind in den vergangenen Jahren extrem stark gesunken. Der ganze Prozess, inklusive Probenentnahme (eine Speichelprobe genügt), Probenaufbereitung, Genotypisierungs-Microarray und Datenanalyse kostet inzwischen nur noch ca. 200 Sfr. pro Teilnehmer. In den meisten GWAS Studien ist die Genotypisierung inzwischen kein erheblicher Kostenfaktor mehr. Andere Aspekte der Studie (Rekrutierung und Betreuung der Teilnehmer, Bestimmung der Phänotypen etc.) sind inzwischen wesentlich kostspieliger.

# rsid	genotype
rs4477212	AA
rs3094315	AT
rs3131972	AG
rs12124819	--
rs11240777	TC
rs6681049	CC
...	
...	**ca. 1’000’000 Zeilen**

Abbildung 3 *Ausschnitt aus der Genotypdatei eines Studienteilnehmers. Jede Zeile repräsentiert einen SNP für den jeweils der rsID und der Genotyp angegeben ist. Je nach Studientyp kann diese Datei zwischen einigen Hunderttausend und einigen Millionen SNPs enthalten. Der Genotyp für alle diese SNPs wird für jeden Teilnehmer benötigt.*

Datenanalyse

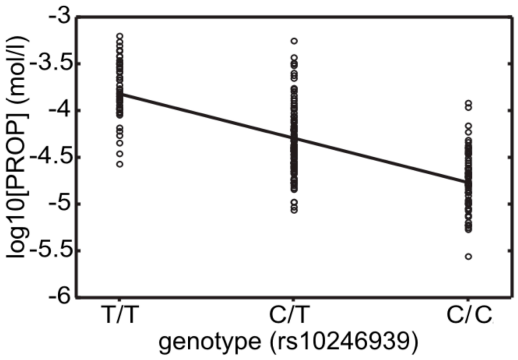

Abbildung 4 *Beispiel eines einzelnen Regressionsplots für einen einzelnen SNP (rs10246939). Die horizontale Achse entspricht dem Genotyp und die vertikale Achse dem Phänotyp. In diesem Fall handelt es sich beim Phänotyp um die minimale Konzentration des Bitterstoffes PROP, den die Teilnehmer in einem Geschmackstest wahrnehmen konnten. Um eine Normalverteilung zu erreichen (siehe nächster Abschnitt, warum das wichtig ist), wird der Phänotyp als Logarithmus dieser Konzentration dargestellt. Die Assoziation von Genotyp und Phänotyp ist hier ungewöhnlich stark ($p < 10^{-50}$).*

Die Grundidee der Datenanalyse in GWAS-Studien ist denkbar einfach. Für jeden der ca. 1'000'000 SNPs wird ein separater Assoziationstest durchgeführt.

Für Studien an quantitativen Phänotypen (z.B. Blutdruck, Cholesterinkonzentration oder Körpergrösse) entspricht dieser Assoziationstest der Bestimmung einer Regressionsgeraden. Bildlich gesprochen, ist dies die Erstellung eines Plots, bei dem für jeden Teilnehmer der Genotyp an diesem SNP gegen den Phänotyp dieses Teilnehmers geplottet wird.

Dann bestimmt man mittels der Methode der kleinsten Quadrate eine Regressionsgerade durch diesen Plot und erhält so die Steigung (β) der Regressionsgeraden und die Standardabweichung der Steigung (σ_β).

Zum Schluss berechnet man aus β, σ_β und n (Anzahl der Teilnehmer) mithilfe der Verteilungsfunktion der Student-t-Verteilung (tcdf) den p-Wert der Assoziation.

the smaller p the better

$$p = 2\ tcdf\left(-\left|\beta/\sigma_\beta\right|; n\right)$$

Dieser p-Wert ist die Wahrscheinlichkeit, mit der man mit zufällig ausgewählten Datenpunkten eine Gerade erhalten würde, die mindestens so stark von der Horizontalen abweicht wie die erhaltene Regressionsgerade.

Aus der Formel ergibt sich, dass eine grössere Steigung der Gerade, eine kleinere Standardabweichung der Steigung und eine grössere Anzahl der Teilnehmer jeweils einen kleineren p-Wert, also eine signifikantere Assoziation, erzeugen.

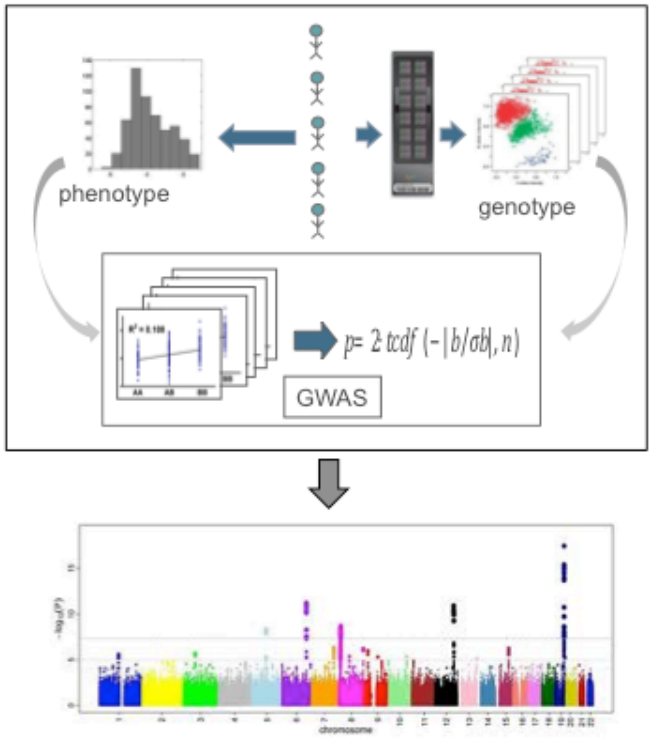

Abbildung 5 *Schematische Darstellung einer GWAS-Studie. Für jeden Teilnehmer misst man den Phänotyp und den Genotyp von ca. 1 Million SNPs. Im Regelfall wird die Genotypinformation mittels Genotypisierungsmicroarrays gemessen. Für jedes der gemessenen SNPs wir dann ein Assoziationstest gegen den Phänotyp durchgeführt und ein p-Wert ermittelt. Diese p-Werte werden dann in einem sogenannten Manhattan-Plot visualisiert.*

Für qualitative Phänotypen (z.B. lockige Haare vs. glatte Haare, oder Herzinfarkt vs. kein Herzinfarkt) verwendet man andere, auf logistischer Regression basierte, statistische Tests. Prinzipiell ist aber auch der Einsatz von anderen statistischen Tests denkbar, solange diese einen p-Wert liefern.

Natürlich werden diese Berechnungen nicht von Hand, sondern von speziell entwickelten Computerprogrammen ausgeführt.

Als primäres Resultat einer GWAS-Studie erhält man somit eine lange Liste von p-Werten. Jeweils einen p-Wert für jeden der ca. 1 Million getesteten SNPs. Diese p-Werte kann man dann in einem sogenannten Manhattan-Plot darstellen, um so auf einen einzigen Blick herauszufinden, ob und wo im Genom besonders starke Assoziationen mit dem Phänotyp auftreten.

Herausforderungen und Annahmen in GWAS-Studien

Im letzten Abschnitt haben wir die Grundidee einer GWAS-Studie besprochen und gelernt wie konzeptionell einfach solche Studien sind. Hier geht es jetzt um eine Reihe von den Dingen, die in der Praxis dazu führen, dass GWAS-Studien dann doch etwas komplizierter ausfallen.

Viele von diesen potentiellen Problemen sind dabei auf eher subtile statistische Effekte zurückzuführen, die man in Einzeltests eventuell vernachlässigen würde.

Da GWAS-Studien aber so viele Einzeltests beinhalten, können selbst kleine Probleme zu relativ grossen systematischen Fehlern führen. Solche Fehler können signifikante Assoziationen vortäuschen und somit zu falschen biologischen Schlussfolgerungen führen.

In diesem Abschnitt werden die wichtigsten potentiellen Probleme besprochen, die in GWAS-Studien auftreten können und wie man versucht, diese Probleme in den Griff zu bekommen.

Normalverteilung und konstante Streuung (Homoskedastizität) von quantitativen Phänotypen

GWAS-Studien basieren auf vielen parallelen Regressionstests. In einer GWAS-Studie müssen also die Daten dieselben Voraussetzungen erfüllen, die auch für einfache Regressionstests gelten. Nur muss man hier, wie oben beschrieben, besonders aufmerksam sein, dass diese Annahmen besonders genau erfüllt sind. Es folgt eine kurze Rekapitulation dieser Annahmen.

Die erste Annahme ist die Normalverteilung der Phänotypen (also der unabhängigen Variablen). Das heisst, nach der Korrektur für den Genotyp und den Covariablen (z.B. Alter, Geschlecht etc.) muss die Verteilung des Phänotyps, also z.B. der Körpergrösse aller Studienteilnehmer, einer Normalverteilung entsprechen. Da der Einfluss von Genotyp und Covariablen auf den Phänotyp oft relativ gering ist, läuft dies darauf hinaus, dass der Phänotyp bereits vor der Korrektur normalverteilt sein sollte. Darüber hinaus sollte die Streuung (die Varianz) der Phänotypen für die unterschiedlichen Genotypen gleich sein (Homoskedastizität). Sind diese Bedingungen nicht erfüllt, sind auch die erhaltenen p-Werte nicht korrekt und können eine höhere Signifikanz vortäuschen.

Abweichungen von der Normalverteilung und auch in gewissem Masse Probleme mit ungleicher Streuung (Heteroskedastizität) lassen sich häufig durch eine Transformation aller Phänotypwerte korrigieren. So könnte man z.B. anstelle des gemessenen Phänotypwertes den Logarithmus dieses Wertes verwenden. Das Alter, bei dem eine bestimmte Krankheit zum ersten Mal auftritt, ist ein Beispiel für einen Phänotyp, der anstelle von normal- oft log-normalverteilt ist. Das gleiche gilt bspw. auch für Metabolitkonzentrationen. In solchen Fällen sollte man dann den Logarithmus des eigentlichen Phänotyps als Wert für den Phänotyp für die GWAS-Analyse verwenden.

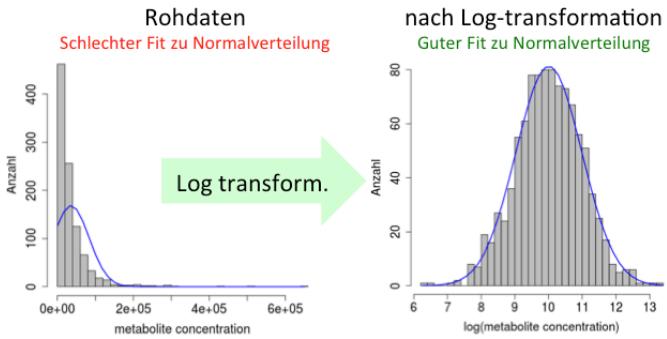

Abbildung 6 *In der Praxis entspricht die Verteilung der Phänotypen häufig nicht einer Normalverteilung. Der Plot links zeigt, wie viele Studienteilnehmer eine bestimmte Konzentration eines Metaboliten in ihrem Blut haben. Die Verteilung entspricht nicht der erwarteten Normalverteilung (der bestmögliche Fit ist als blaue Linie angezeigt). Der rechte Plot zeigt dieselbe Analyse, verwendet aber statt der Konzentration des Metaboliten den Logarithmus dieser Konzentration. Der Log-transformierte Phänotyp entspricht nahezu perfekt einer Normalverteilung und ist daher besser als Input-Phänotyp für eine GWAS-Studie geeignet.*

Im Prinzip lässt sich jede Verteilung durch eine entsprechende Transformation in eine Normalverteilung zwingen (z.B. durch Rangnormalisierung). Wenn aber derart handgreifliche Methoden notwendig sind um eine Normalverteilung zu erreichen, ist es vielleicht sinnvoll darüber nachzudenken, ob es nicht eine andere Messgrösse gibt, die denselben biologischen Prozess erfasst und dessen Verteilung näher an der gewünschten Normalverteilung liegt.

Phänotyp-Rauschen & Covariablen

Für manche Phänotypen ist die experimentelle Bestimmung denkbar einfach und auch präzise. Die Körpergrösse ist z.B. ein solcher Phänotyp. Für diese Art von Phänotypen muss man sich wenig Gedanken über die Bestimmung des Phänotypen (Phänotypisierung) machen.

Andererseits gibt es auch Phänotypen, die schwer zu messen sind (z.B. die Messung von bestimmten Metaboliten in der Leber oder die Schwere der Depression eines Studienteilnehmers). Für solche Phänotypen macht es sehr viel Sinn sich intensive Gedanken über die optimale Phänotypisierung zu machen, denn die Qualität, mit der die Phänotypen gemessen wurden, ist einer der entscheidenden Faktoren für den Erfolg einer GWAS-Studie. In

Fällen in denen der Phänotyp zeitlich stark schwankt oder in denen die Messungen sehr ungenau sind, könnte man z.B. den Phänotyp durch Mittelung über mehrere Messungen präzisieren. Andererseits ist es aber auch sinnvoll, über alternative Phänotypen nachzudenken, die auch Einblick in denselben biologischen Prozess bieten, aber vielleicht einfacher zu messen sind.

Im Regelfall kann man davon ausgehen, dass der beobachtete Phänotyp nicht nur durch den Genotyp beeinflusst wird, sondern auch durch eine Vielzahl von anderen Faktoren. Bei der Konzentration eines Lebermetaboliten könnten diese anderen Faktoren z.B. der Zeitpunkt der Probenentnahme, die Ernährungsgewohnheiten oder das Geschlecht des Studienteilnehmers sein.

Von der Perspektive der Genotyp-Phänotyp-Assoziation betrachtet, generieren diese anderen, den Phänotyp ebenfalls beeinflussenden Faktoren, ein Rauschen, das wir so weit als möglich unterdrücken wollen. Neben dem Phänotyp selber, misst man also in der Regel auch eine Reihe von Faktoren, die den Phänotyp beeinflussen könnten. Im GWAS-Kontext nennt man diese Faktoren auch Covariablen. Bei der Auswahl der Covariablen, die für einen bestimmten Phänotyp relevant sind, ist ein fundiertes Verständnis des untersuchten biologischen Prozesses besonders wertvoll.

Unabhängigkeit der Einzelbeobachtungen

Idealerweise sollten in einer GWAS-Studie sowohl Genotypen und Phänotypen, als auch alle anderen Eigenschaften der Teilnehmer, unabhängig von einander sein.

Eine systematische Korrelation zwischen dem Phänotyp unterschiedlicher Teilnehmer sollte einzig dadurch entstehen, dass sie an einem, den Phänotyp beeinflussenden, Locus die gleichen genetischen Variationen besitzen.

Der Grund für diese Bedingung liegt darin, dass eine grössere Anzahl von unabhängigen Einzelbeobachtungen (sprich Teilnehmern) zu einer grösseren Signifikanz der Assoziation führt. Wenn Korrelationen zwischen den einzelnen Beobachtungen existieren, reduziert sich also die effektive Zahl der unabhängigen Beobachtungen und die Signifikanz der Assoziation ist eventuell wesentlich geringer als man basierend auf der Teilnehmerzahl annehmen würde.

Das folgende Gedankenexperiment illustriert diesen Punkt. Man nehme eine Studie mit 200 Teilnehmern. Aus dem letzten Abschnitt wissen wir, dass wenn die anderen Parameter (β und σ_β) gleichbleiben, eine grössere Teilnehmerzahl einen kleineren p-Wert ergibt.

Wir könnten also auf die Idee kommen, einfach alle Teilnehmer zweimal an der Studie teilnehmen zu lassen. Dadurch steigt n auf 400 und wir erhalten durch diese „Doppelnutzung“ unserer Teilnehmer für alle SNPs einen wesentlich geringeren p-Wert, also eine höhere Signifikanz unserer Assoziationen.

Natürlich ist die eigentliche Information, die wir in dieser Studie zur Verfügung haben, exakt gleichgeblieben. Der so erhalten p-Wert überschätzt also die Signifikanz.

Man kann sich diese Doppelnutzung von 200 Teilnehmern als einen Extremfall einer Studie von 400 Teilnehmern vorstellen, bei der sowohl die Genotypen, als auch die Phänotypen sehr stark miteinander korreliert sind. Hier fällt die Überschätzung der Signifikanz also besonders stark aus. Wenn die Korrelation geringer ausfällt, fällt sicherlich auch die Überschätzung der Signifikanz geringer aus, das Grundproblem bleibt jedoch bestehen.

Indirekte Assoziationen, Bevölkerungsstruktur (engl. *population stratification)*

Wie wir wissen, entstehen genetische Variationen durch Zufallsprozesse und sind nicht das Resultat phänotypischer Variationen. Die Gensequenz eines Menschen verändert sich also **nicht,** weil er/sie dick, gross oder rothaarig ist oder weil er/sie Diabetes hat!

Wenn wir also eine Assoziation zwischen Genotyp und Phänotyp beobachten, sollte die Kausalitätsrichtung eigentlich klar sein: der Genotyp beeinflusst den Phänotyp und nicht umgekehrt.

Dies ist grundsätzlich korrekt, aber es besteht eine weitere Möglichkeit. Ein dritter Faktor könnte sowohl mit dem Genotyp, als auch dem Phänotyp, korreliert sein und so indirekt eine Assoziation zwischen Genotyp und Phänotyp verursachen, obwohl kein kausaler Zusammenhang zwischen Genotyp und Phänotyp besteht.

In der Praxis sind solche indirekten Assoziationen oft auf *population stratification* zurückzuführen. *Population stratification* beschreibt dabei die Tatsache, dass die Teilnehmer genetisch nicht aus einer einzigen homogenen Population, sondern aus distinkten Unterpopulationen stammen. In diesem Sinne kann *population stratification* auch als ein spezieller Fall von Genotypkorrelation gesehen werden.

Solche Unterpopulationen entstehen immer dann, wenn die zufällige Durchmischung der Genotypen innerhalb einer Population über längere Zeiträume verhindert wird, z.B. durch geographische oder kulturelle Barrieren. Dieselben geographischen und kulturellen Barrieren bedingen darüber hinaus häufig auch unterschiedliche Lebensweisen bezüglich Ernährung, körperlicher Aktivität, medizinischer Versorgung, etc. Die Präsenz von Unterpopulationen in einer Studie kann also sehr leicht zu indirekten Assoziationen zwischen Genotyp und Phänotyp führen.

Idealerweise sollte die Teilnehmerpopulation einer Studie daher so ausgewählt werden, dass keine *population*

stratification vorliegt. Praktisch lässt sich das aber kaum erreichen. Man ist daher in den letzten Jahren von einer Vermeidungs- zu einer Korrekturstrategie übergegangen. Dabei hat sich besonders die sogenannte EigenStrat-Methode als sehr effizient erwiesen. Bei dieser Methode werden die Genome der Teilnehmer einer mathematischen Anaylse, die sich *principal component analysis* (PCA) nennt, unterzogen. Ohne in Details zugehen sei hier gesagt, dass eine PCA die sehr hochdimensionalen genetischen Koordinaten der Teilnehmer (jeder gemessene SNP entspricht einer Dimension) auf eine minimale Anzahl von Achsen „projiziert". Dies erlaubt es, den Hauptanteil der genetischen Unterschiede zwischen den Teilnehmern durch eine kleine Anzahl von Koordinaten (Eigenwerten) entlang bestimmter Achsen (Eigenvektoren) darzustellen. Diese so gewonnenen Koordinaten werden dann in der GWAS-Analyse als Covariablen verwendet um *Populationsstruktur*-Effekte zu korrigieren.

GWAS-Auswertung: die Schlüsselparameter

Während die ersten GWAS-Studien zum Teil noch recht unterschiedliche Analyseansätze verfolgten, hat sich inzwischen ein Standardprozess herauskristallisiert, der für die überwiegende Mehrzahl der GWAS-Studien verwendet wird.

1) Auswahl eines Phänotyps der so leicht und präzise zu messen ist, wie möglich und zugleich einen guten Einblick in den zu untersuchenden biologischen Prozess erlaubt.
2) Abschätzung der Erblichkeit des Phänotyps und der Anzahl der Teilnehmer, die für eine genomweit signifikante Assoziation notwendig ist (nicht immer möglich).
3) Rekrutierung der Teilnehmer, Messung des Phänotyps und der Covariablen.
4) Bestimmung des Genotyps durch Genotypisierungs-Microarrays.
5) Qualitätskontrolle der SNPs.
6) Imputation von zusätzlichen SNPs (optional).
7) Bestimmung der Bevölkerungsstruktur durch *principal component analysis*.
8) Bestimmung der Covariablen, die den Phänotyp signifikant beeinflussen.
9) Korrektur des Phänotypen für signifikante Covariablen
10) GWAS Analyse
11) Korrektur der erhaltenen p-Werte durch *genomic control*
12) Generierung von QQ- und Manhattan Plots

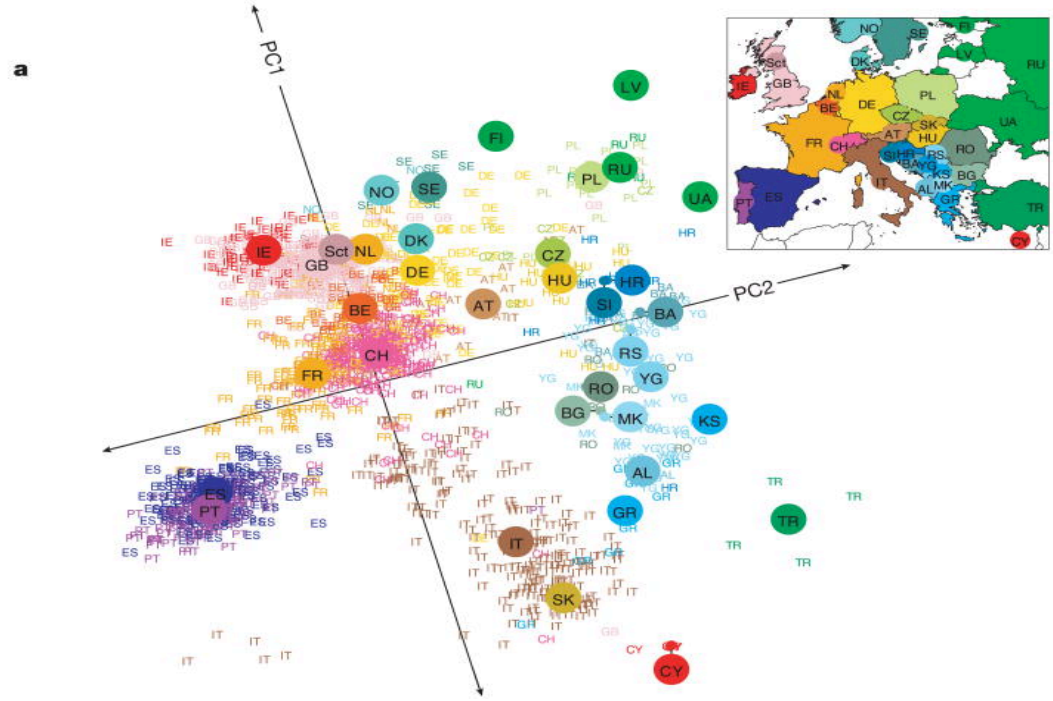

Abbildung 7 *Principal component analysis (PCA) des Genoms ist eine effektive Methode zur Erfassung der Bevölkerungsstruktur (engl. population stratification). Die Abbildung zeigt einen Scatter-Plot der beiden ersten principal components der Genome einer Gruppe von Europäern, wobei der Herkunftsort der Grosseltern der Teilnehmer durch ein farbiges Buchstabenkürzel angegeben ist (es wurden nur Teilnehmer untersucht bei denen alle vier Grosseltern aus demselben Land stammen). Wie zu sehen ist, entspricht die Position in diesem Plot weitestgehend den Koordinaten ihrer Herkunft. Die Genomdaten selbst können also genutzt werden um Bevölkerungsstruktur zu erkennen und zu korrigieren (aus Novembre et al., Nature 2008).*

Unterschiede zwischen den verschiedenen Studien liegen inzwischen hauptsächlich in der Art und Weise, wie der Phänotyp aufbereitet wird.

Diese standardisierte Vorgehensweise macht es relativ einfach, die Resultate verschiedener Studien miteinander zu vergleichen und die Qualität der beobachteten Assoziationen zu beurteilen. Dabei sind einige Parameter und Plots von besonderer Bedeutung und werden hier im Detail besprochen.

p-Wert, Bonferroni-Korrektur und genomweite Signifikanz

Wie bereits beschrieben, besteht eine GWAS-Studie aus einer langen Serie von Einzeltests. In den Biowissenschaften werden einzelne statistische Tests häufig dann als statistisch signifikant angesehen, wenn die Wahrscheinlichkeit ein so extremes Testresultat per Zufall zu erhalten weniger als 1% ist (also $p < 0.01$).

Wir wissen aber, dass wenn man einen Zufallstest oft genug wiederholt, früher oder später auch Resultate beobachtet werden, die in einem einzelnen Test als sehr unwahrscheinlich angesehen würden (Beispiel: Wenn man lange genug würfelt, wirft man irgendwann einmal drei Sechsen hintereinander).

Bei einer Million getesteter SNPs erwartet man also schon aus purem Zufall, dass eine ganze Reihe dieser Assoziationstest einen p-Wert von kleiner als 0.01 liefern.

Bei Mehrfachtests müssen wir also unsere Signifikanzkriterien entsprechend anpassen.

Die konzeptionell einfachste Methode um für solche Mehrfachtests zu kompensieren, ist die sogenannte Bonferroni-Korrektur. Dabei wird der Wert unter welchem p-Werte als signifikant betrachtet werden durch die Anzahl der durchgeführten Tests geteilt. Auf die 1 Million Einzeltests einer GWAS-Studie angewendet, müsste also ein Assoziationstest für einen der getesteten SNPs einen p-Wert von $< 0.01/1'000'000 = 0.00000001 = 10^{-8}$ erreichen um als signifikant zu gelten. Die Hürde für Signifikanz liegt bei einer GWAS-Studie also sehr hoch. Es sind daher in der Regel Studien mit sehr vielen Teilnehmern notwendig, um diese hohe statistische Signifikanz zu erreichen.

Die Bonferroni-Korrektur geht dabei davon aus, dass die Einzeltests einer Testreihe voneinander völlig unabhängig sind. Wie wir aber gesehen haben, sind die Genotypen von benachbarten SNPs durch *linkage disequilibirum* miteinander korreliert. Es bestehen also nicht wirklich eine Million unabhängige Chancen per Zufall um einen p-Wert von 10^{-8} zu erreichen. Mit anderen Worten: die Bonferroni Korrektur schiesst über das Ziel hinaus.

Man verwendet daher inzwischen einen Signifikanz-Schwellenwert (*cutoff*) von $10^{-7.5}$ für GWAS-Studien, selbst wenn wesentlich mehr als eine Million SNPs verwendet werden.

Wenn bereits einer der SNPs diesen *cutoff* überschritten hat, gilt ausserdem der nächste SNP schon als signifikant assoziiert, wenn er einen p-Wert von $10^{-7.5}/2$ hat und so weiter.

Manhattan-Plots: Ein Bild sagt mehr als tausend Worte

Ein Manhattan-Plot ist die wohl informativste und leichtverständlichste Art, die Ergebnisse einer GWAS-Studie darzustellen. Dafür wird jeder einzelne SNP in einem Scatter-Plot eingetragen, wobei die x-Achse die Position

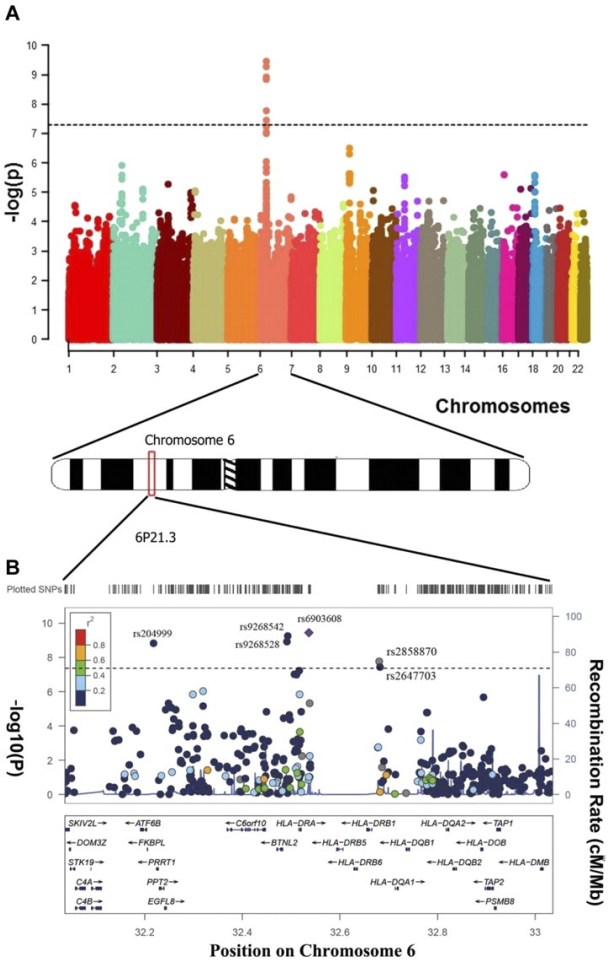

Abbildung 8 *(A) Beispiel eines typischen Manhattan-Plots. Durch die grosse Anzahl der abgebildeten SNPs verschmelzen die Punkte im unteren Teil des Plots zu einer einheitlichen Fläche. Die gepunktete Linie zeigt den Signifikanz-Schwellenwert an. (B) Abschnitt des Manhattan-Plots mit starker Vergrösserung entlang der Genomachse (der horizontalen Achse). Man nennt solche Plots „lokale Manhattan-Plots", wobei unter dem Plot oft die Positionen der in dieser Genomregion befindlichen Gene angezeigt sind. Der hier gezeigte Plot stammt aus einer GWAS-Studie an einer Blutkrebsart, dem Hodgkin Lymphoma (Cozen et al. blood, vo.l 119, 2011).*

im Genom (d.h. Chromosomennummer und Basenpaarposition) und die y-Achse den p-Wert der Assoziation mit dem Phänotyp abbildet.

Die y-Achse ist in diesen Plots normalerweise logarithmisch dargestellt. Genauer gesagt wird die Stärke der Assoziation als –log10 des p-Werts dargestellt (also p = 0.0000001 entspricht –log(p)=7). Je höher der Punkt, desto stärker die Assoziation. Ein Manhattan-Plot enthält also normalerweise mehrere hunderttausend, vielleicht sogar einige Millionen einzelne Datenpunkte, so dass diese Datenpunkte im unteren Teil des Plots miteinander verschmelzen. Der optische Gesamteindruck dieser Plots erinnert ein wenig an die Skyline einer Grossstadt, wodurch diese Plots ihren Namen erhalten haben. Oft, aber nicht immer, zeigt eine horizontale Linie im Plot den *cutoff* für statistische Signifikanz an.

QQ-Plots – Qualitätskontrolle auf einen Blick

Basierend auf unserem bisherigen Wissen erwarten wir, dass im Regelfall nur eine Handvoll Gene den untersuchten Phänotyp beeinflussen. Für alle anderen Gene und die darin liegenden SNPs, also die weit überwiegende Mehrzahl der in einer GWAS-Studie getesteten SNPs, sollte die Assoziation mit dem Phänotyp den Gesetzen des Zufalls folgen. Wir können diese Tatsache zur Qualitätskontrolle nutzen. Mit anderen Worten: Wenn die p-Werte der weit überwiegenden Mehrzahl der SNPs nicht der erwarteten Zufallsverteilung entsprechen, ist dies ein ziemlich sicheres Zeichen, dass in der Analyse etwas falsch gelaufen ist. Insbesondere wollen wir vermeiden, dass Probleme zu einer Inflation der p-Werte, also zu fälschlichen Assoziationen und somit falschen Schlussfolgerungen, führen.

Dabei gilt für Zufallsprozesse generell, dass sich die beobachtete und die erwartete Verteilung umso mehr gleichen sollten, je grösser die Anzahl der Beobachtungen ist. Im Fall von GWAS-Studien mit fast einer Million Beobachtungen erwartet man also eine nahezu perfekte Übereinstimmung mit der erwarteten Zufallsverteilung und man kann so selbst kleine Abweichungen leicht erkennen.

Die erwarte Zufallsverteilung bei einer GWAS-Studie mit 1 Million SNPs ist, dass 1 SNP aus 1'000'000 einen p-Wert von 1/1'000'000 = 10^{-6} erreicht. Der zweitstärkste assoziierte SNP hätte einen erwarteten p-Wert von 2/1'000'000, der hundertstärkste einen p-Wert von 100/1'000'000 = 10^{-4} , usw. bis zum schwächsten assoziierten SNP, für den ein p-Wert von 1'000'000/1'000'000 = 1 = 10^{-0} erwartet wird.

Um zu testen, ob diese Verteilung in der Tat beobachtet wird, verwendet man die sogenannten QQ-Plots (für Quantile-Quantile; siehe Abbildung 9 & Abbildung 10). Dabei sortiert man die experimentell beobachteten p-Werte aller SNPs der Grösse nach und macht dies ebenfalls für den Satz von p-Werten der oben beschriebenen Zufallsverteilung. Dann plottet man die so entstandenen Wertepaare in einem Scatter-Plot. Wie beim Manhattan-Plot verwendet man dabei nicht den p-Wert selber, sondern den $-\log_{10}$ des p-Wertes.

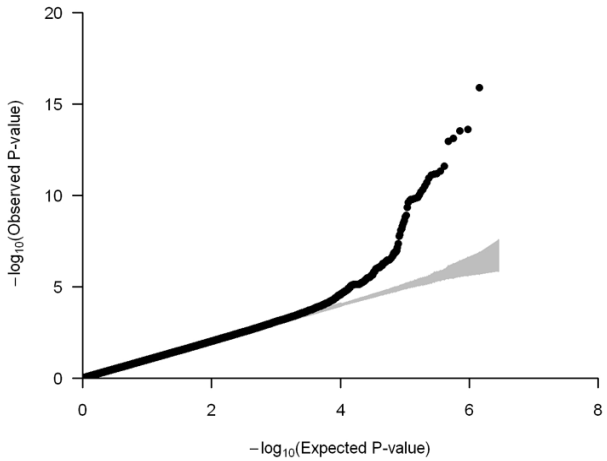

Abbildung 9 *Beispiel eines mustergültigen QQ-Plots. Die Datenpunkte im linken unteren Bereich des Plots liegen perfekt auf der Diagonalen des Plots, sie folgen also perfekt der erwarteten Zufallsverteilung. Im rechten oberen Teil des Plots finden sich einige Dutzend SNPs, bei denen die beobachteten p-Werte deutlich über den erwarteten p-Werten liegen. Diese SNPs sind also signifikant assoziiert. Der graue Bereich zeigt den Bereich des Plots in dem man per Zufall assoziierte SNPs erwartet.*

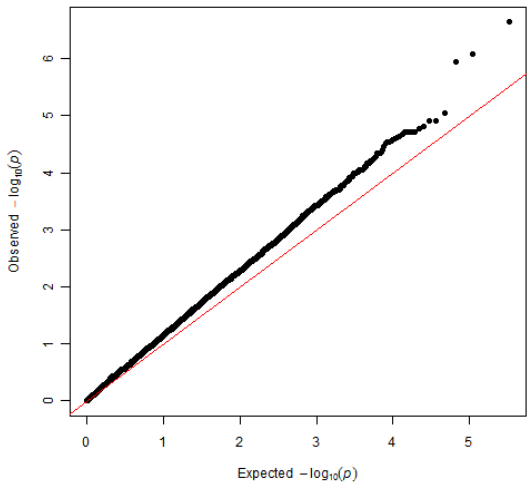

Abbildung 10 *Ein QQ-Plot, der auf p-Wert Inflation hindeutet. In diesem QQ-Plot liegen alle SNPs (selbst die im unteren linken Bereich) oberhalb der Diagonalen. Dies deutet auf systematische Probleme hin. Die SNPs, die in diesem Plot oben rechts liegen, haben vielleicht einen p-Wert, der alleine genommen eine signifikante Assoziation anzeigen würde. Angesichts der systematischen p-Wert Inflation, die sich im Rest des Plots zeigt ist diese Signifikanz aber fragwürdig.*

Falls die experimentell beobachtete p-Wert-Verteilung nun der per Zufall erwarteten Verteilung entspricht, sollten die Wertepaare alle auf der Diagonalen des Plots liegen. Punkte, die stark über der Diagonale liegen deuten dann auf signifikante Assoziationen hin. Sind aber alle Punkte systematisch ober- oder unterhalb der Diagonalen, deutet dies auf systematische Probleme bei der Analyse hin (siehe Abbildung 10).

Genomic control und λ: kleinere Probleme lassen sich leicht nachträglich korrigieren

QQ-Plots sind selten so perfekt wie in Abbildung 9 gezeigt. Es ist dabei durchaus üblich, dass man leichte systematische Abweichungen des Plots beobachtet. Die einer GWAS-Studie zugrundeliegenden Annahmen sind eben selten alle perfekt erfüllt. Oft ist es dabei unrealistisch alle Gründe für diese leichten Abweichungen zu finden und zu korrigieren. Stattdessen führt man nachträglich eine sogenannte *genomic control* Korrektur durch, die eine leichte systematische Über- oder Unterschätzung der p-Werte korrigiert. Dabei wird der Parameter λ bestimmt, der die Abweichung des QQ-Plots von der Diagonalen im unteren Bereich misst – dort, wo wir mit an Sicherheit grenzender Wahrscheinlichkeit eine Zufallsverteilung erwarten. Diese Korrektur wird dann auf alle p-Werte angewendet, also auch auf den Bereich des QQ-plots in dem wir eventuell signifikante Assoziationen erwarten. *Genomic control* sollte dabei nur als eine Art „Feinschliff" eingesetzt werden und sollte keinesfalls benutzt werden, um gröbere Probleme „unter den Teppich zu kehren". Der λ-Parameter ist daher auch ein gutes Qualitätsmass für eine GWAS-Studie. Ein λ von 1 bedeutet, dass keinerlei Korrektur nötig war. Werte zwischen 0.95-1.05 gelten als durchaus akzeptabel. Grössere Abweichungen weisen oft auf ernsthafte Probleme hin, die nicht mehr durch *genomic control* korrigiert werden sollten und eine Suche nach den tatsächlichen Ursachen wird nötig.

ß und r^2 (für quantitative Phänotypen)

In der Interpretation von GWAS-Studien fällt oft sehr viel Augenmerk auf den p-Wert und die diversen Plots der p-Wert-Verteilung (siehe oben). Die Parameter β, die Steigungen der Regressionsgeraden, und r^2, der Anteil der phänotypischen Varianz, der durch die Regressionsgerade „erklärt" wird, erhalten dabei oft vergleichsweise wenig Aufmerksamkeit. Dies aber ganz zu Unrecht:

β ist eigentlich der biologisch relevanteste Parameter einer GWAS-Studie, weil er anzeigt, wie stark und in welche Richtung ein SNP mit dem Phänotyp assoziiert ist. Der r^2-Parameter (auch *explained variance* genannt) beschreibt dabei, welcher Anteil der phänotypischen Varianz insgesamt durch den Genotyp an diesem SNP „erklärt" wird. Der Wert von r^2 kann von 0 (der SNP-Genotyp erklärt überhaupt nichts) bis 1 (der SNP-Genotyp alleine erklärt die gesamte Phänotypvarianz) reichen. Ein r^2 von 1 würde dabei einer den Mendel'schen Gesetzen gehorchenden Beziehung von Genotyp und Phänotyp entsprechen. Da ein substantieller Teil der Phänotypvarianz meist auf Phänotyprauschen (z.B. Ungenauigkeit in der Messung des Phänotyps) zurückgeführt werden kann, spricht man aber schon bei r^2-Werten von über ca. 50 % von Mendel'schen Assoziationen.

Odds ratio (für qualitative Phänotypen)

Für qualitative Phänotypen übernimmt der sogenannte *odds ratio* (OR) die Rolle, die β und r^2 für quantitative Phänotypen einnehmen. Der *odds ratio* zeigt an, wie der Genotyp das Verhältnis von Studienteilnehmern, die den Phänotyp aufweisen, zu denen die den Phänotyp nicht aufzeigen (z.B. an Diabetes erkrankt oder gesund) beeinflusst. Abbildung 11 zeigt ein Beispiel.

In der Regel wird bei qualitativen Phänotypen die Assoziation zwischen Genotyp und Phänotyp durch logistische Regression untersucht, wobei der logistische Regressionskoeffizient β_L bestimmt wird. Dabei gilt dann $OR = e^{\beta L}$.

	+	-
T/T	254	182
T/C	345	312
C/C	167	214

$O_{(TT,TC)} = (254 + 345)/(182+312) = 1.21$
$O_{(CC)} = (167)/(214) = 0.78$
$OR = O_{(TT,TC)} / O_{(CC)} = 1.55$

Abbildung 11 *Beispiel für die Berechnung eines odds ratio. Hier gehen wir bei der Berechnung davon aus, dass das Merkmal, welches von Allel T bestimmt wird, dominant ist. Der odds ratio zeigt an, dass in dieser Studie Träger des T-Allels eineinhalbmal häufiger den Phänotyp aufweisen als Teilnehmer, die homozygot für das C-Allel sind. Wenn der Phänotyp z.B. Diabeteserkrankung wäre, würde dies bedeuten, dass das Vorhandensein des T-Allels ein moderat erhöhtes Krankheitsrisiko bedeutet.*

Zusammenhang zwischen Teilnehmerzahl (n), erklärter Varianz (r^2), statistischer Signifikanz (p) und phänotypischem Rauschen

Die Parameter n, r^2 und p stehen über die kumulative Verteilungsfunktion der Student-t-Verteilung in einem direkten mathematischen Zusammenhang. Dabei gilt:

$$p = 2 \cdot \mathrm{tcdf}\left(-\sqrt{n\frac{r^2}{1-r^2}}, n\right)$$

Die Formel ist ein wenig komplex, aber es sollte ersichtlich sein, dass n und r^2 sich gegenseitig kompensieren (Hinweis: für kleine r^2 gilt die Approximation $r^2/(1\text{-}r^2) \approx r^2$).

Konkret bedeutet dies, dass man denselben p-Wert entweder dadurch erreichen kann, dass man in einer Studie mit wenigen Teilnehmern einen SNP mit grossem Einfluss auf den Phänotyp hat, oder dass man in einer sehr grossen Studie einen SNP hat, der wenig von der phänotypischen Varianz erklärt. Basierend auf der obigen Formel lässt sich der in Abbildung 12 gezeigte Graph berechnen, der dieses Verhältnis quantitativ beschreibt. Wenn man z.B. in einer GWAS-Studie nach einem SNP sucht, der mindesten 30% der phänotypischen Varianz erklärt, dann braucht man mindesten 100 Teilnehmer in der Studie damit ein solcher SNP einen genomweit signifikanten p-Wert von $10^{-7.5}$ erreicht.

Ein anderer Aspekt, der schon im vorhergehenden Abschnitt angesprochen wurde, aber trotzdem oft übersehen wird, ist, dass die von einem SNP erklärte phänotypische Varianz nicht ausschliesslich von biologischen Faktoren abhängt, sondern auch davon, wie genau der Phänotyp gemessen wurde. Formal ausgedrückt setzt sich die totale phänotypische Varianz aus einzelnen Beiträgen zusammen:

$$var_{total} = var_{umwelt} + var_{genetisch} + var_{rauschen} + \cdots$$

Die Ungenauigkeiten bei der Messung des Phänotyps tragen also auch zur totalen phänotypischen Varianz bei. Wenn die totale Varianz also grösser wird, sinkt entsprechend der Anteil der totalen Varianz, der durch den genetischen Einfluss eines SNPs erklärt wird. Bei der Planung einer GWAS-Studie kann man also seine Chancen auf eine genomweit signifikante Assoziation nicht nur dadurch verbessern, dass man eine grössere Anzahl von Teilnehmern rekrutiert, sondern auch dadurch, dass man den Phänotypen geschickt auswählt und besonders akkurat misst.

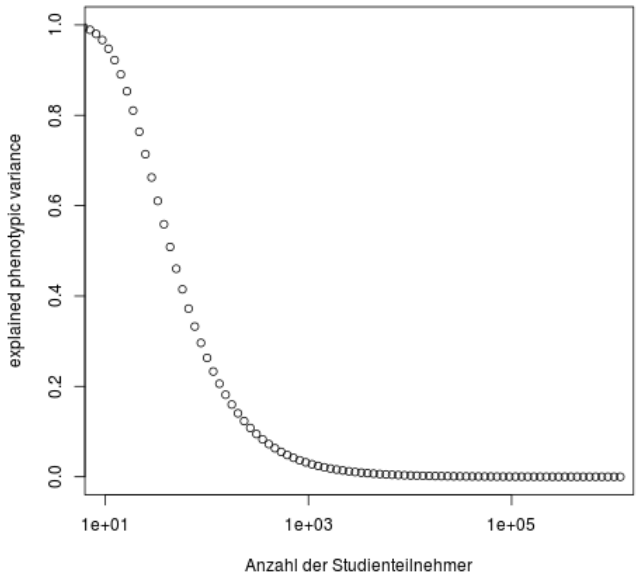

Abbildung 12 *Der Graph zeigt den Anteil der phänotypischen Varianz (r^2), den ein SNP erklären muss, um in einer Studie mit einer bestimmten Anzahl von Teilnehmern eine genomweit signifikante Assoziation ($p=10^{-7.5}$) zu generieren.*

GWAS-Replikation und Follow-Up

Gehen wir einmal davon aus, dass wir eine GWAS-Studie erfolgreich abgeschlossen haben. Das heisst wir haben SNPs gefunden, die signifikant mit unserem getesteten Phänotyp assoziiert sind. Was machen wir nun?

Replikation

GWAS-Studien sind derart komplex und die Effekte auf den Phänotyp oft so gering, dass sich trotz der besten Kontrollen immer noch irgendwo ein systematischer Fehler einschleichen kann. Mit anderen Worten, die beobachtete Assoziation könnte ein Artefakt sein.

Wenn es sich bei der beobachteten Assoziation also nicht um einen ungeheuer starken Effekt (r^2 > 30%) handelt und/oder die durch die assoziierten SNPs implizierten Gene nicht sofort einen plausiblen biologischen Mechanismus für den beobachteten Phänotyp nahelegen, dann verlangt man inzwischen, dass die Resultate einer GWAS-Studien in einer zweiten, unabhängigen Studie repliziert werden.

Idealerweise benutzt die Replikationsstudie dabei eine andere Genotypisierungsplattform (z.B. DNA-Sequenzierung anstatt SNP-Chips), rekrutiert die Teilnehmer aus einer anderen Bevölkerungsgruppe und misst den Phänotyp auf eine etwas andere Art als die ursprüngliche Studie. Man versucht also alle möglichen Quellen von systematischen Fehlern auszuschliessen.

Replikationsstudien können dabei oft mit deutlich weniger Teilnehmern durchgeführt werden, als die ursprüngliche Studie. Dies liegt daran, dass man in der Replikationsstudie ja jetzt ein Kandidatengen bzw. einen Kandidaten-SNP hat und man die Assoziation des Phänotyps mit nur diesem SNP testet. Mit anderen Worten ist eine Bonferroni-Korrektur ist nicht mehr nötig.

Wenn die Replikation erfolgreich ist, kann man sich ziemlich sicher sein, dass es sich bei der Assoziation nicht um ein Artefakt handelt. Dann stellt sich die Frage nach dem Mechanismus, der den assoziierten SNP mit der phänotypischen Variation verknüpft.

Kausal- oder Proxy-SNP?

Ein kausaler SNP ist der SNP, der den Phänotyp mechanistisch beeinflusst. Anders ausgedrückt, würde man im Genom eines Menschen nur dieses einzelne Basenpaar austauschen würde sich der entsprechende Phänotyp dieses Menschen verändern. Das entsprechende Experiment darf man an einem Menschen natürlich nicht durchführen!

Leider sind die SNPs, die man in einer GWAS-Studien identifiziert nur in den seltensten Fällen die kausalen SNPs, sondern meistens sind es sogenannte Proxy-SNPs. Solche Proxy-SNPs sind SNPs, die mit den kausalen SNPs bzw. den kausalen genetischen Variationen (z.B. CNVs oder Indels) in starkem LD stehen, also mit hoher Wahrscheinlichkeit gemeinsam vererbt werden. Assoziationsdaten alleine können nicht klären, ob ein SNP ein kausaler oder ein Proxy-SNP ist.

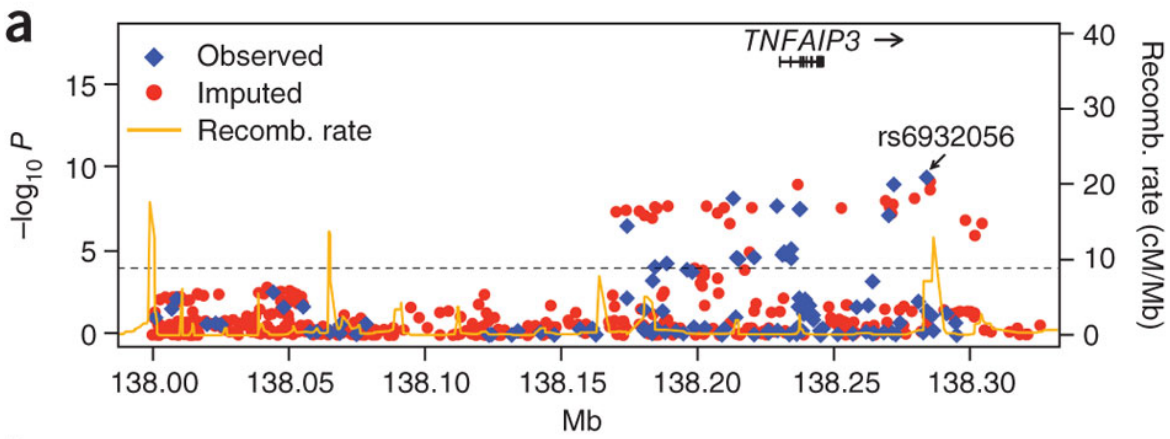

Abbildung 13 *Lokaler Manhattanplot aus einer GWAS-Studie an der Autoimmunkrankheit lupus erythematodes (Adrianto et al. Nature Genetics 2010). Zwei Dutzend SNPs, die über eine Region von ca. 100'000 bp verteilt sind, sind ungefähr gleich stark mit dem Phänotyp assoziiert. Dies bedeutet nicht, dass alle diese SNPs einen mechanistischen Einfluss auf den Phänotypen haben. Vielmehr ist es wahrscheinlich, dass diese SNPs in einen LD Block fallen, also mit sehr hoher Wahrscheinlichkeit gemeinsam vererbt werden. Die Variation die wirklich einen mechanistischen Einfluss auf den Phänotyp hat, befindet sich irgendwo in diesem LD Block.*