In den Manhattenplots sieht es so aus als ob es an manchen SNP Positionen mehrere Punkte gibt. Wie ist das möglich, ich dachte für jeden SNP gibt es nur einen p-Wert. Wofür stehen die anderen Punkte?

Sie haben recht, jeder Punkt in einem Manhattan Plot repräsentiert den p-Wert der Assoziation zwischen dem untersuchten Phänotyp und einem SNP der an einer bestimmten Position im Genom liegt. Der Eindruck, dass mehrere Punkte an derselben horizontalen Position, also am selben SNP liegen ist nur eine „optische Illusion". In Manhattenplots werden so viele SNPs dargestellt, dass benachbarte SNPs so aussehen als ob Sie an derselben Position im Genom liegen. Zoomt man in diese Plots hinein (siehe z.B. der local Manhatten Plot auf Slide 22 der Vorlesung) so sieht man, dass diese anscheinend übereinanderliegenden Punkte in Wirklichkeit lateral gegeneinander verschoben sind, also verschiedenen SNPs zugeordnet sind.

Was genau passiert mathematisch bei der Korrektur des Phänotyps für eine Co-Variable?

Aus der linearen Regression zwischen dem Phänotyp und der Co-Variable erhält man eine Regressionsgerade (Steigung und Y-Achsen Abschnitt). Diese Regressionsgerade beschreibt wie sich der Phänotyp im Mittel mit der Co-Variablen verändert. In unserem Beispiel der Regression des Alters gegen die wahrgenommene Intensität, gibt die Steigung der Regressionsgeraden an um wie viele Intensitätpunkte die Bewertung im Durchschnitt für jedes Lebensjahr abnimmt. Basierend auf dieser Information und der Annahme, dass der Effekt des Alterns auf alle Probanden gleich wirkt, kann man nun für jeden Probanden die Intensität berechnen die dieser Proband bei einem andere Alter wahrnehmen würde. Dazu addiert man für ältere Probanden zu der tatsächlichen Wahrnehmungsintensität einen altersproportionalen Korrekturfaktor.
Umgekehrt könnte man für jüngere Personen den entsprechenden Korrekturfaktor abziehen. Beide Vorgehensweisen sind korrekt und ergeben bei der anschliessenden Assoziation des Genotyps mit dem korrigierten Phänotyp denselben p-Wert.

Mathematisch wird der korrigierte Phänotyp Y_{corr} aus dem unkorrigierten Phänotyp Y_{uncorr} der Steigung b der Regressionsgerade und dem Wert C der Co-Variable wie folgt berechnet.

$$Y_{corr} = Y_{uncorr} - b \times C$$

Habe ich das richtig verstanden, dass GWAS nur möglich sind, wenn man mehrere Personen vergleicht? (Ansonsten könnte man z.B. gar keinen p-Wert finden)

Ja, dass haben Sie richtig verstanden. Wenn an einer Studie nur eine Person teilnehmen würde hätte der Regressionsplot zwischen Genotyp und Phänotyp nur einen Punkt und es wäre nicht möglich eine eindeutige Regressionsgerade zu bestimmen und somit gäbe es auch keine p-Wert.

Was heisst, dass man für den Genotyp korrigieren muss in einer GWAS?

Mir ist nicht klar worauf die Frage abzielt. Gerne nochmal per Email nachfragen!

Manhatten Plot (Abbildung 8): Entspricht jeder Punkt einer Person? Wie kann man einen p-Wert einer einzelnen Person bestimmen?"

Nein, die Punkte in einem Manhattenplot entsprechen jeweils einem SNP und geben den p-Wert der Regression zwischen dem Phänotyp und dem Genotyp (an diesem SNP) aller Probanden an.

Wo liegt der Zusammenhang zwischen Linkage Disequilibrium und GWAS?
Der Zusammenhang zwischen Linkage Disequilibrium ist und GWAS ist mir noch nicht ganz klar.
Ich verstehe noch nicht ganz den Zusammenhang zwischen linkage disequilibrium und GWAS.
Im Skript zu den SNP habe ich die letzte Abbildung zu den Haplotypen und dem LD-Block noch nicht ganz verstanden und würde eine Erklärung im Plenum dazu begrüssen.

Die Erläuterungen in der Vorlesung haben hoffentlich bereits geholfen diesen Punkt zu klären. LD bedeutet, dass die Genotypen von benachbarten SNPs mit einander korreliert sind. Ist einer der SNPs im LD Block mit dem Phänotyp korreliert so sind es die anderen SNPs in diesem LD Block auch. Es reicht also aus, wenn man pro LD Block eine Handvoll SNPs genotypisiert. Das reduziert Arbeitsaufwand und Kosten einer GWAS Studie. Dies bedeutet aber auch, dass alle SNPs, die im selben LD Block liegen wie der SNP, der den Phänotyp wirklich verursacht (kausaler SNP), auch mit dem Phänotyp korreliert sind. Für den Experimentator ist es dabei nicht klar welcher dieser SNPs der kausale SNP ist.

Woher stammen die Daten, mit denen GWAS durchgeführt werden? Ich kann mir auch vorstellen, dass man für ein vollständiges Bild über die vorhandenen SNPs im menschlichen Genom Daten aus verschiedenen Kulturen etc. benötigt, wie ist das z.B. mit Personen aus sehr isolierten Kulturen oder aus Ländern wie z.B. Nordkorea oder auch China, die es sicher nicht zu oberst auf ihrer Prioritätenliste haben, Daten für GWAS zu sammeln? Und wie vertrauenswürdig sind verwendete Daten? In der Systembiologievorlesung wurde letzte Woche erwähnt, dass es dem Dozenten schon passiert ist, dass er mit Daten gearbeitet hat, die von einem Studenten gesammelt wurde und dieser im Nachhinein nicht mehr wusste, welche seine Trainingsdaten waren. Wie geht man in solchen Situationen vor?
Die Genomanalysen werden meistens in grossen Servicelabors generiert die durch Automatisierung und entsprechende Arbeitsprozesse sehr zuverlässig und reproduzierbar arbeiten. Man sammelt also Blut oder Speichelproben von den Teilnehmern und schickt diese an ein solches Servicelabor. Die Probenröhrchen sind zumeist mit Barcode versehen und der ganze Prozess wird anhand dieser Barcodes nachverfolgt. In diesem Prozess sind Fehler sehr selten.
Fehler passieren wohl eher bei der Ausgabe der Proberöhrchen, also wenn ein bestimmter Barcode einer Person zugeordnet wird.
Um solche Fehler zu finden führt man bei der Datenanalyse eine Qualitätskontrolle durch. Z.B. sollten bei Proben die von Männern stammen alle SNPs auf dem X-Chromosom homozygot sein. Ist das für mehrere Proben nicht der Fall, weiss man, dass man irgendwo in der Datensammlungskette ein Problem hat. Viele Studien führen zusätzlich zu der Phänotypisierung für die eigentliche Forschungsfrage auch noch eine Phänotypisierung für einen Phänotyp (z.B. Bitterkeit von PROP) durch, von dem man weiss, dass er eine starke Assoziation zeigen wird. Ist diese Assoziation dann schwächer (i.e. der r-square ist geringer als erwartet) so wurde vermutlich irgendwo schlampig gearbeitet wurde (Proben

vertauscht, ungenaue Phänotypisierung etc.) und man weiss, dass die Daten unzuverlässig sind.

Eine Frage zum Quiz Frage 6: Es könnte doch leider sein, dass die Bestimmung des Phänotyps, also eine nicht perfekte Charakterisierung auch zu einem höheren r-square wert führen kann und es somit eine bessere Korrelation gibt, was leider zu einer Verfälschung mit besserem; Ergebnis führen könnte.
Ja das ist theoretisch möglich, aber bei der Anzahl an Probanden in einer typischen GWAS Studie ist dies extrem unwahrscheinlich. Genau darum geht auch bei der Berechnung des p-Werts der Assoziation zwischen Genotyp und Phänotyp. Hier fragt man wie wahrscheinlich es ist, dass eine so starke Assoziation, wie sie an einem bestimmten SNP beobachtet wurde, per Zufall auftreten könnte.

Im Handout zu den GWAS zum Menschen würden mich noch mehr Hintergrundwissen zu der mathematischen Formel der population stratification interessieren.
Diese Analyse basiert auf einer PCA (Principal Component Analysis). Ganz grob beschrieben ist dies eine Technik mit der man in hochdimensionalen Daten diejenigen Achsen findet entlang denen die grösste Varianz der Datenpunkte besteht.
Diese Analyse basiert darauf, dass man den Genotyp einer Person als einen Punkt in einem n-dimensionalen Koordinatensystem beschreiben kann, wobei n die Anzahl der SNPs im Datensatz ist. Hat eine Person z.B. 2 Kopien des Variantallels an einem bestimmten SNP, so liegt die Koordinate die diese Person beschreibt zwei „Ticks“ entlang der entsprechenden Achse. Einen Raum mit 1 Million Dimension kann man aber nicht visualisieren. Man sucht also nach derjenigen Achse in diesem Raum entlang derer die Varianz zwischen den Datenpunkten (i.e. den einzelnen Personen) so gross wie möglich ist. Diese Achse ist die erste Principal Component (PC1). Dann sucht man nach einer zweiten Achse welche orthogonal zu PC1 ist und entlang der, der nächstgrösste Grad an Varianz besteht. Dies ist dann PC2 usw. Auf diese Weise erhält man ein neues Koordinatensystem mit den Axen PC1, PC2 etc. in dem sich der Genotyp einer Person darstellen lässt. Der Vorteil dieses Koordinatensystems ist, dass eine Projektion auf die PC1 / PC2 Ebene die maximale Varianz der Daten zeigt, also zwei Punkte die in diesem Plot weit auseinander liegen gehören zu Leuten die genetisch sehr unterschiedlich sind. Falls diese recht simplifizierte Erklärung nicht tiefgenug geht, werden Sie auf Wikipedia oder in vielen Mathematik Lehrbüchern oder auf Youtube fündig. Mathematisch betrachtet ist eine PCA übrigens ein Eigenvektor/Eigenwertproblem für dessen Lösung sehr effiziente Algorithmen zur Verfügung stehen.

Eine PCA wird meinem Verständnis nach ausgeführt, um sekundäre, nicht genetisch bedingte Varianz zu erkennen und aus den Daten filtern zu können. Also Varianz die nur ungewollt korreliert. Können aber nicht gerade z.B. ein geographischer Parameter mit der Häufigkeit eines SNPs korrelieren und dieser Effekt dann durch die PCA wieder herausgerechnet werden, wodurch es schwieriger ist, signifikante Resultate zu bekommen?
Ja, das haben Sie sehr gut erkannt, im Prinzip kann durch PCA Korrektur von Population Stratification ein Teil des Signals in einer GWAS „wegkorrigiert“ werden. In der Praxis ist dies aber ein relativ vernachlässigbares Problem. Der Grund dafür ist, dass die meisten SNPs nur relativ schwach mit den Principal Components der Population Stratification korrelieren.

Ist das nicht der Fall und die Korrelation zwischen dem kausalen SNP und einem PC ist stark, hat man in der GWAS grosse Probleme. Denn dann korrelieren ja alle SNPs die zu diesen PCs beitragen auch indirekt mit dem Phänotyp. Man würde also eine grosse Anzahl von SNPs finden die alle mit dem Phänotyp korrelieren obwohl kein funktionaler Zusammenhang besteht. Das ist ja genau die Situation die man durch eine Population Stratification Korrektur vermeiden will. In der Praxis bedeutet dies, dass eine Population, in der ein grosser Teil der phänotypischen Variation durch einen der PCs erklärt wird, nicht wirklich gut für eine GWAS Studie an diesem Phänotyp geeignet ist.

Können sie in der Vorlesung bitte noch einmal auf Proxy SNPs eingehen? also ensteht das LD nur durch räumliche Anordnung auf dem Chromosom oder gibt es andere Gründe.
Das sollte nach der Vorlesung eigentlich klar sein. Wenn noch nicht gerne noch einmal emailen.

Hat nun, wenn z.B. ein Phänotyp von mehreren SNPs beeinflusst wird und zwei dieser SNPs miteinander korrelieren (nahe Loci, oft gemeinsam vererbt) und somit beide (obwohl nur eines tatsächlich Auswirkung hat) im manhatten plot auftauchen das Verfahren versagt?
Nein, das GWAS Verfahren hat in dem Fall geleistet was es leisten kann. Es hat unter 1 Million SNPs eine sehr kleine Gruppe von SNPs identifiziert in deren unmittelbarer Umgebung sich das (oder die) kausalen SNPs befinden.

Ich bin etwas verwirrt. Im Skript „Genetische Vielfalt beim Menschen (SNPs)“ wird im Abschnitt „SNPs starten als private Punktvariante“ erwähnt, dass ein Elternpaar 2 Kinder hat und dass einer der beiden Eltern für eine autosomale Variante 1 heterozygot ist. Geht man davon aus, dass der andere Elternteil homozygot für eine andere Variante 2 ist, sollte die Chance, dass beide Kinder die Variante 1 erben ebenfalls bei 25 % liegen. Ist die Wahrscheinlichkeit, dass die Variante 1 in der Population erhalten bleibt nicht gleich gross wie die Wahrscheinlichkeit dass sie verschwindet?
Sie haben recht, im Durchschnitt bleibt die Allelfrequenz eines Allels in einer Population über die Generationen hinweg konstant. Aber durch Zufall kann sie nach oben und unten fluktuieren. Wenn die Allelfrequenz sehr gering ist, (also z.B. nur eine Kopie des Allels in der ganzen Population vorliegt) besteht eine relativ grosse Wahrscheinlichkeit, dass die Allelfrequenz durch eine solche Fluktuation einmal auf null abfällt. Damit ist dieses Allel dann verschwunden und kann nur durch eine weitere, statistisch sehr unwahrscheinliche, Neumutation wieder in die Population zurückkommen. Die Wahrscheinlichkeit, dass ein Allel durch eine solche Fluktuation aus der Population verschwindet ist umso grösser je geringer die Allelfrequenz ist. Private Mutation, die nur einmal in der ganzen Population vertreten sind, haben also das höchstmögliche Risiko aus der Population zu verschwinden.

Ich habe nicht vollständig verstanden, welche Aussagen sich über den Parameter Beta machen lassen.
Der Parameter beta ist die Steigung der Regressionsgeraden zwischen Genotyp und Phänotyp. Er gibt an wie sich der Phänotypwert eines „durchschnittlichen“ Probanden mit zunehmender Anzahl (0, 1, oder 2) der Kopien des Variantenalles an einem SNP verändert.

Folgendes Beispiel ist vielleicht hilfreich. Nehmen wir an der untersuchte Phänotyp ist die Körpergrösse in cm. Nehmen wir weiter an wir haben einen SNP dessen „Referenzallel“ G und dessen Variantenallel T ist. Beträgt beta für die Regressionsgerade für diesen SNP -0,2 so bedeutet dies, dass Probanden mit Genotyp TT an diesem SNP im Durchschnitt 0.4cm kleiner sind Probanden mit Genotyp GG. Der absolute Wert von beta sagt also aus wie gross der Effekt eines SNPs auf einen quantitativen Phänotyp ist und das Vorzeichen sagt aus welches der Allele mit einem höheren und welches mit einem niedrigeren Phänotypwert assoziiert ist.

Was genau versteht man unter Rauschen eines Phänotyps?
Rauschen in diesem Zusammenhang sind Variation im gemessenen Phänotyp, die durch nicht-genetische Faktoren hervorgerufen werden. Untersucht man als Phänotyp z.B. das Körpergewicht der Probanden so würden die Verwendung von mehreren Waagen, die nicht untereinander kalibriet sind, eine Quelle von Phänotyprauschen darstellen, weil sie eine Variation der gemessenen Körpergewichte verursachen, die aber nichts mit Genetik zu tun hat. Eine andere Quelle von Phänotyprauschen in diesem Beispiel wäre, wenn man einige Probanden nach und andere Probanden vor dem Frühstück wiegt.

Skript: Genetische Vielfalt beim Menschen: Focus auf SNPs, S.3 „[..] selbst Erwachsene Lactase produzieren und daher Milch verdauen können. In den meisten Regionen der Welt haben Träger dieser Variante eher einen kleinen Fitness Nachteil“
Wieso ist die „Lactose-Toleranz“ ein kleiner Fitness Nachteil?
Die einsetzende Laktoseintoleranz begrenzt die Dauer der Stillzeit wodurch die Mutter ihre biologischen Ressourcen in eine erneute Schwangerschaft investieren kann und so ihre Gene an mehr Nachfahren weitergeben kann. Fehlt dieser „Anreiz“ hat der Säugling die Tendenz länger zu stillen und so mehr der Ressourcen der Mutter für sich zu beanspruchen als er benötigt. Die Interessen von Mutter, Vater und Kind sind aus einer, für die menschliche Vorpflanzung inzwischen kaum mehr relevanten, rein biologischen Sicht durchaus unterschiedlich. Dies führt übrigens zu recht komplexen genetischen und epigenetischen Effekten.

Wenn man ein GWAS gut durchgeführt wird, wie kann man wissen welche spezifische Basenpaaren verursachen das untersuchte Phänotyp?
Mich würde interessieren, wie man die Unterscheidung von Proxy vs. Causal SNP im Menschen Experimentell bestimmt.
Wie geht man vor, wenn man mithilfe einer GWAS ein SNP gefunden hat (und sie auch repliziert wurde), um den Mechanismus zu finden, welchen das SNP beeinflusst?

Eine Analyse der genetischen Daten kann diese Frage im Regelfall nicht lösen. Es sind also biochemische, zellbiologische etc. Untersuchungen gefragt. Welche Experimente geeignet sind diese Frage zu klären hängt stark vom Phänotypen und vom SNP ab. Fallen die SNPs z.B. in die proteinkodierende Sequenz eines Enzymgens könnte man die beiden daraus resultierenden Proteinvarianten herstellen und im Reagenzglass testen ob einer der Variationen die Funktion des Proteins verändert. Wenn dann auch noch bekannt ist, dass eine verminderte Funktion dieses Enzyms den beobachteten Phänotyp verursacht, so hat man eine recht überzeugende Kausalkette zwischen SNP und Phänotyp und kann davon ausgehen, dass dieser SNP in der Tat kausal ist.

Für Experiment an Modelorganismen bietet sich inzwischen die CRISPR/Cas9 Technologie an, mit der man gezielt die einzelnen Kandidaten SNPs in einen Organismus einführen und dann den resultierenden Phänotyp beobachten kann. Verursacht einer der SNPs den untersuchten Phänotyp ist er wohl der kausale SNP.

Die SNP einer genomweiten Assoziationsstudie sind genetische Abweichungen vom Referenzgenom. Was wird als Referenzgenom definiert?
Das Referenzgenom wurde vom Human Genome Consortium festgelegt. Das ist eine relativ arbiträre Definition die als Standard verwendet wird. Dabei ist es für die GWAS egal welches Allel eines SNPs als Variante und welches als Referenz betrachtet wird. Der p-Wert der Assoziation bleibt der gleiche, es verändert sich nur das Vorzeichen von beta.

Sind die erwähnten Haplotypen auch dafür verantwortlich das z.B. blaue Augen und blondes Haar oft gemeinsam vererbt werden?
Nein, soweit ich weiss liegt das häufige gemeinsame Auftreten von blauen Augen und blonden Haaren nicht daran, dass sich die verantwortlichen SNPs in LD befinden, sondern daran, dass sowohl blaue Augen als auch blonde Haare in den selben Populationen einen evolutionären Vorteil hatten und die entsprechenden Varianten dadurch in diesen Populationen besonders häufig auftreten. Das häufigen gemeinsame Auftreten von blonden Haaren und blauen Augen erklärt sich also eher durch einen Population Structure Effekt als durch einen LD Effekt.

Was ist genomic control (bei Punkt 11.: Korrektur der erhaltenen p-Werte durch genomic control), bzw. wie wird sie genau durchgeführt?
Weiter unten im Skript (Seite 9) sollten Sie fündig werden.

Gibt es Nachteile bei der „handgreiflichen“ Transformation zur Normalverteilung?
Der vermutlich grösste Nachteil besteht in dem Effekt von Messfehlern auf die transformierten Phänotypen. Bei einer linearen Regression gehen wir davon aus, dass die Messfehler des Phänotyps normal verteilt und für jeden Probanden gleich gross sind (diese Annahme ist ohnehin schon recht „mutig“!). Führt man nun eine Transformation durch, werden zufällige Messfehler des Phänotyps nach oben und unten nicht gleichmässig von der Transformation beeinflusst. Führt man z.B. eine logarithmische Transformation durch, hat ein zufälliger Messfehler nach unten systematisch einen grösseren Einfluss als ein zufälliger Messfehler nach oben. (100+-20 wird nach der Transformation zu 2 +0.0791 -0.0969.) Je grösser die Messfehler und je drastischer die Transformation desto gravierende ist dieser Effekt und desto weniger sind die einer GWAS zugrundeliegenden Annahmen gerechfertigt. Das Endresultat sind dann häufig eine grosse Anzahl von Assoziationen die später nicht reproduzierbar sind.

Das halbieren des cutoffs nach dem überschreiten durch den ersten SNP verstehe ich nicht. Auf Seite 7 steht: 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.

Wieso ist das so?

Dieses Vorgehen wird als Benjamini-Hochberg Verfahren bezeichnet. Dieses Verfahren ist nicht spezifisch für GWAS Studien, sondern findet seine Anwendung überall dort wo die Signifikanz von Mehrfachtests bewertet werden muss.

Generell will man bei statistischen Assoziationstests möglichst viele Assoziationen finden bei denen ein wirklicher Zusammenhang zwischen den zwei Variablen besteht. Gleichzeitig will man möglichst all diejenigen Assoziationen verwerfen, die nur auf Zufall beruhen. Legt man extrem strenge Kriterien an die Signifikanz an, verwirft man zu viele „echte" Assoziationen. Sind die Kriterien zu locker werden zu viele „falsche" Assoziationen als signifikant angesehen. Die Herausforderung ist immer eine gute Balance zwischen diesen beiden Arten von Fehlern zu finden.

In diesem Kontext beschäftigt sich das Benjamini-Hochberg Verfahren mit dem speziellen Fall in dem mehrere Assoziationen aus einer Testserie nahe an der Signifikanzschwelle liegen. Der Grundgedanke hinter diesem Verfahren ist, dass es in einer Serie von Einzeltests unwahrscheinlicher ist, dass mehrere dieser Einzeltests eine bestimmte Signifikanzschwelle überschreiten als das nur einer der Einzeltests diese selbe Schwelle überschreitet. Hat man also in einer Testserie schon eine Assoziation gefunden die es über die sehr hohe, durch die Bonferroni Korrektur gesetzte, Signifikanzschwelle geschafft hat, ist es recht unwahrscheinlich, dass in dieser Testserie, per Zufall, auch noch eine zweite Assoziation dieser Schwelle überhaupt nur nahekommt. Beobachtet man also eine solche zweite Assoziation nahe an der Signifikanzschwelle ist die Wahrscheinlichkeit hoch, dass diese zweite Assoziation nicht per Zufall so stark ist, sondern eine „echte" Assoziation repräsentiert. Es würde also Sinn machen für die zweite Assoziation eine etwas geringere Signifikanzschwelle anzulegen und für eine dritte Assoziation eine noch geringere usw.. Dabei will man die Signifikanzschwelle aber auch nicht so tief legen, dass zu viele „falsche" Assoziationen durchkommen. Den Überlegungen von Benjamini und Hochberg zufolge liefert die um dem Faktor ½, 1/3 etc. angepassten Signifikanzschwellen für die zweitstärkste, drittstärkste etc. Assoziation die optimale Balance.

Wenn die im Skript erläuterten Berechnungen nochmals verdeutlicht werden könnten, wäre dies sehr hilfreich.

Sie haben ja alle relevanten Berechnungen nun in der Besprechung einmal selber durchgeführt. Falls es noch Unklarheiten gibt, gerne noch einmal per Email oder über das Kursforum melden.

Bei den 12 Punkten zum Standardprozess in GWAS: Was bedeutet Punkt 5) „Qualitätskontrolle der SNPs" ? Wie macht man diese Qualitätskontrolle?

Die Qualitätskontrolle der SNPs basiert normalerweise auf der Entfernung von SNPs:
- bei denen die Verteilung der Allele nicht der Hardy-Weinberg Verteilung entspricht.
- bei denen das seltenere der zwei Allele zu selten auftritt (cutoff ist oft 5%)
- deren Bestimmung bei zu vielen Probanden fehlgeschlagen ist (cutoff ist oft 5%)
- SNPs bei denen die Qualität des Fluoreszenssignals auf dem Chip zu schwach war (cutoff abhängig von Genotypisierungsinstrument)

Wieso werden die kausalen SNPs nicht entdeckt?
Wurde in der Vorlesung besprochen. Wenn trotzdem noch nicht klar gerne im Forum oder per Email nachfragen.

Wie viel Statistik muss man für die Prüfung können?
Gäbe es eine Möglichkeit, dass das noch genau definiert wird?
Ich habe den Abschnitt mit den verschiedenen Parametern noch nicht ganz verstanden.
Eins der Lernziele lautet: „Sie können sich anhand einiger Schlüsselparameter (p-Wert, n, lambda und r^2) und Abbildungen (Manhattenplot, QQ-plot) einen Eindruck von der Qualität und Relevanz einer in der Primärliteratur veröffentlichten GWAS Studie machen."
In der Besprechung haben Sie gesehen das es im Prinzip dabei nur um lineare Regressionen geht. Es hilft vielleicht noch einmal die entsprechenden Materialien aus der Statistik Vorlesung herauszusuchen und sich damit zusammen noch einmal das Material aus der Besprechung anzuschauen. Bei spezifischen Fragen gerne noch einmal im Forum oder per Email nachfragen.

Wie für seltene Krankheiten üblich, haben unterschiedliche Tay-Sachs-Patienten in der Regel unterschiedliche Mutationen.
Würde es nicht eher Sinn machen, wenn es nur eine/wenige Mutationen sind, welche eine seltene Krankheit ausmachen, da die Krankheit ja nur selten ist. Bei vielen möglichen Mutationen würde sie ja häufiger auftreten.
Bei seltenen genetischen Krankheiten ist es zumeist so, dass in verschiedenen Familien in denen dieselbe Krankheit auftritt diese Krankheit aber durch unterschiedliche genetische Variationen verursacht wird. Das einzige was die Variationen der verschiedenen Familien gemeinsam haben ist, dass sie dieselbe biologische Funktion verändern, also z.B. dasselbe Enzym inaktivieren, und so dieselbe Krankheit hervorrufen.
Es sind dabei jeweils nur eine Handvoll Individuen, die die jeweilige Mutation tragen. Diese Mutationen sind in der Regel evolutionär noch sehr jung, haben sich also noch nicht in der Gesamtbevölkerung ausgebreitet (vermutlich werden sie dies auch nie tun, weil sie ja eine schwere Krankheit verursachen, also einen Selektionsnachteil darstellen).
Weitverbreitete Krankheiten (z.B. Diabetes oder hohes Cholesterin) werden hingegen oft von SNPs beeinflusst die vor sehr langer Zeit einmal entstanden sind, sich dann über tausende Generationen in der Bevölkerung ausgebreitet haben und nun im Genom von Milliarden von Menschen zu finden sind. Erst jetzt, wo die entsprechende Überernährung besteht, werden diese SNPs aber zu einem „Problem".

Ich verstehe den Manhattan-Plot immer noch nicht ganz. Wäre froh um eine Erklärung, vor allem auch mit Bezug auf den P-Wert.
Ich komme nicht ganz draus, wie der p-Wert den Zusammenhang des Genotyps und des Phänotyps erklären sollte
Hoffe das ist nach der Besprechung jetzt klar. Wenn nicht gerne nochmal melden.

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 zur toxischen Anreicherung des Substrats. Von welchem Substrat ist hier die Rede?

In der Bioinformatik Vorlesung zu Beginn des Kurses haben Sie gelernt wo man diese Art von Information nachschauen kann. Setzen Sie das Gelernte hier doch direkt einmal um. Falls es nicht klappt und Sie es wirklich gerne wissen wollen, gerne noch einmal melden.

Ist die Genauigkeit von GWAS für Studien an quantitativen und qualitativen Phänotypen vergleichbar?

Generell ist es, wenn möglich, vorteilhaft für GWAS Studien eine quantitativen Phänotypen zu verwenden. Man findet so typischerweise bereits mit deutlich weniger Probanden signifikante Assoziationen. Hat man aber wirklich nur einen qualitativen Phänotyp zur Verfügung kann man mit einer entsprechend grösseren Anzahl Probanden genauso zuverlässige Assoziationen erhalten.