Aus dem Institut für Medizinische Genetik der Medizinischen
Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

# Strukturvarianten im menschlichen Genom

zur Erlangung des akademischen Grades
Doctor medicinae (Dr. med.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Fabian Grubert

aus Berlin

Gutachter: 1. Prof. Dr. med. S. Mundlos

2. Prof. Dr. med. O. Rieß

3. Prof. Dr. med. H.-H. Ropers

**Datum der Promotion: 19.11.2010**

## *Widmung*

Ich widme diese Arbeit meinen Eltern

# *Inhaltsverzeichnis*

# 1 Einleitung

## 1.1 Allgemeines

Was ist die genetische Grundlage für phänotypische Variation? Wie unterscheiden sich verschiedene Individuen hinsichtlich ihrer Erbinformation? Welchen Einfluss haben Unterschiede in unserem Erbmaterial darauf, wie wir auf verschiedene Umwelteinflüsse reagieren? Was bestimmt unsere Prädisposition für Krankheiten? Mit diesen Fragen beschäftigt sich die Humangenetik seit ihren Anfängen. Mit Einführung der Zytogenetik konnte man beobachten, dass mikroskopisch sichtbare Veränderungen der Karyotypen, entweder in Form von Verlust oder Dazugewinn von chromosomalen Segmenten oder gar ganzer Chromosomen, die Grundlage für viele genetische Erkrankungen, wie z. B. Trisomie 21 [1], darstellen. Fortschreitende Technologie hat es ermöglicht, submikroskopische Veränderungen bis hin zu Unterschieden auf der Ebene der Nukleinsäuren festzustellen. Ein klassisches Beispiel ist die Punktmutation im Beta-Globin-Gen als Ursache der Sichelzellanämie [2]. Je mehr Veränderungen gefunden wurden, desto klarer wurde, dass nicht alle Variationen im genetischen Code mit „Krankheit“ assoziiert sein müssen. Vielmehr entdeckte man, dass sich die DNA-Sequenz von zwei „normalen“ Individuen in Millionen von Basen unterscheiden kann. Diese Basenunterschiede sind über das Genom verteilt und lassen sich im Schnitt etwa alle 800 Basenpaare (bp) finden. Wenn eine solche Variante an einer Stelle eine Häufigkeit von größer als 1 Prozent in der Bevölkerung hat, spricht man von einem Single-Nucleotide-Polymorphism (SNP). Bis vor wenigen Jahren nahm man an, dass diese SNPs die Haupt-Grundlage für genetische Unterschiede zwischen Individuen darstellen und sich auch die meisten genetischen Erkrankungen auf sie zurückführen lassen und ihnen eine entscheidende Rolle in der Pharmakogenetik zukommt. Deshalb haben akademische Forschung und Industrie sich verstärkt darauf konzentriert, alle SNPs in menschlichen Populationen zu finden und ihre relative Häufigkeit zu kartieren [3-5]. Mittlerweile sind mehr als 10 Millionen SNPs bekannt, und neben ihrer direkten Bedeutung als Strukturvariante (SV) sind sie essentiell als Marker in genomweiten Assoziationsstudien (GWAS), die es möglich machen, auch nach Kandidatenregionen bzw. -genen in komplexen Erkrankung zu suchen.

Im Rahmen der Initiative, die gesamte Erbinformation des Menschen zu sequenzieren und zu kartieren (Humanes Genom Projekt) [6], wurde eine detaillierte Referenzsequenz des menschlichen Genoms erstellt. Mit Abschluss des Projektes wurde angenommen, dass sich zwei gesunde Individuen nahezu ausschließlich in ihren SNPs unterscheiden, abgesehen von sehr seltenen Varianten, die große Abschnitte von Chromosomen betrafen. Weitere vergleichende Studien führten zu einem überraschendem Ergebnis: Neben den SNPs gibt es eine weitere Klasse von weitverbreiteten Polymorphismen, die Deletionen, Insertionen, Inversionen (siehe Abbildung 1) und Translokationen mit einer Größe von ca. 1 Kb bis mehreren Megabasen umfasst. Diese Art der Variation kann ganze Gruppen von Genen betreffen, und mittlerweile konnte gezeigt werden, dass sie nicht nur Einfluss auf unseren Phänotyp hat, sondern auch mit Prädisposition für Krankheit assoziiert sein kann.

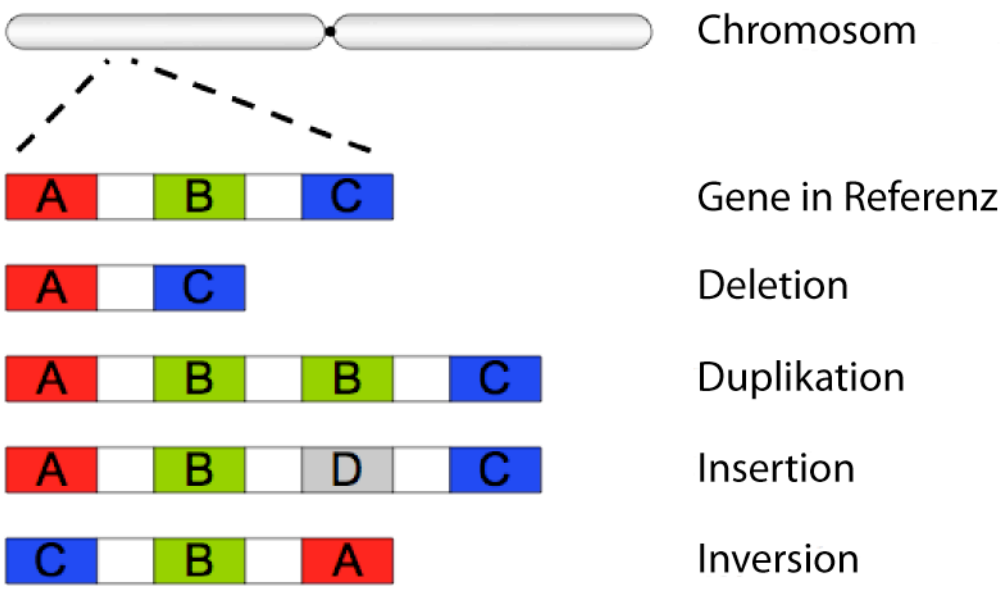



**Abbildung 1: Mögliche Strukturvarianten**
Verschiedene Klassen von Polymorphismen: Deletion, Duplikation, Insertion und Inversion.

## 1.2 Kopienzahlveränderung und Chromosomenaberrationen in genetischen Erkrankungen

Neben dem Austausch einzelner Basen in der DNA und daraus folgender Veränderung der kodierten Proteinstruktur, wie z. B. im Falle der Sichelzellanämie

[7], können auch Kopienzahlveränderungen (engl.: Copy number variation (CNV)) die Grundlage für Erbkrankheiten sein. Das häufigste Beispiel hierfür in der Klinischen Genetik ist die Trisomie 21 (Down-Syndrom). Hierbei betrifft die Kopienzahlveränderung große Teile oder - wie in 95% der Fälle - ein ganzes Chromosom 21, das in betroffenen Individuen in dreifacher Kopie im Chromosomensatz vorhanden ist. Es ist also nicht ein einzelnes Gen oder eine Gruppe von Genen in ihrer Struktur verändert, sondern alle Gene auf Chromosom 21 liegen statt in zweifacher in dreifacher Ausführung in der Zelle vor. Die Chromsomenaberration kann auch wie im Falle des Katzenschrei-Syndroms (Cri-du-chat-Syndrom) in einem teilweisen Verlust eines Chromosomes bestehen, was dazu führt, dass der betroffene Teil nur noch in einfacher Ausführung auf dem homologen Schwesterchromosom liegt. Diese Erkrankungen beruhen also auf Deletion oder Duplikation bestimmter Chromosomenabschnitte. Das Spektrum dieser Chromosomenaberrationen erstreckt sich vom mikroskopisch sichtbaren bis in den submikroskopischen Bereich. Hier, bei den Mikrodeletions- bzw. Duplikationssyndromen, wird das Krankheitsbild durch die Deletion oder Duplikation von Chromosomenabschnitten kleiner als 2 Mb verursacht. Als Beispiel sei die Charcot-Marie-Tooth (CMT) Neuropathie Typ 1A genannt. Hier führt eine 1,4 Mb-große Duplikation [8] auf Chromosom 17 u. a. zur Kopienzahlzunahme von PMP22 (peripheral myelin protein 22), einem Dosis-sensitiven Gen, das ein Protein im Myelin des Peripheren Nervensystems codiert. Mit verbesserten Detektionsmethoden, wie z. B. der hochauflösenden vergleichenden genomischen Hybridisierung mithilfe von DNS-Rastern (engl.: Array comparative genomic hybridizations (aCGH)), konnten zahlreiche neue Mikrodeletions- bzw. Duplikationssyndrome identifiziert werden, und die weiter steigende Zahl unterstreicht ihre Bedeutung und möglichen Einfluss auf bisher noch nicht vollständig geklärte genetische Syndrome.

## 1.3 Die Wirkung von Strukturvarianten und Kopienzahlveränderungen auf den Phänotyp ist komplex

Der Effekt von Änderungen in der Kopienzahl von Genen ist vielgestaltig und nicht jede Abweichung vom „normalen“ Genotyp hat notwendigerweise Folgen für den Organismus. Das Spektrum der möglichen Auswirkungen reicht von nicht nachweisbaren Normvarianten über Einfluss auf komplexe, multifaktorielle

Erkrankungen bis hin zu Mikro-Deletions- bzw. Mikro-Duplikations-Syndromen, bei denen die Kopienzahlveränderung der alleinige krankheitsverursachende Faktor ist. Manche Duplikationen oder Deletionen, die die Kopienzahl bestimmter Gene beeinflussen, können sich als euchromatische Varianten präsentieren, die keinen nachweisbaren Einfluss auf den Phänotyp haben [9, 10]. Eine solche Variante kann aber auch von Vorteil für den betroffenen Organismus sein und zur Adaption des Organismus an seine Umwelt beitragen, wie im Falle des Amylase-Gens (AMY1). Kopienzahlveränderungen desselben führen über proportional veränderte Amylase-Konzentrationen im Speichel zur Anpassung des Organismus an die verfügbare Nahrung, hier an den Anteil an Stärke in der Diät [11]. Eine weitere Gruppe von Genen, die weitreichende, ausgesprochene Variabilität in Kopienzahlen zeigt, ist die Klasse der Olfaktorischen Rezeptor-Gene, für die hunderte verschiedener CNVs gefunden wurden [12, 13]. Dies ist ein weiterer Hinweis auf die Bedeutung der CNVs für die Evolution des Menschen.

Manche Kopienzahlveränderungen können die Empfänglichkeit für bestimmte Erkrankungen beeinflussen ohne selber hinreichende Ursache zu sein. So konnte beispielsweise eine Korrelation zwischen der Kopienzahl von MIP-1alphaP, welches CCL3L1 kodiert, und erhöhter Empfänglichkeit für HIV gezeigt werden [14]. CCL3L1 ist ein HIV-1-supprimierendes Chemokin, das als Ligand am HIV-Rezeptor CCR5 wirkt. Weitere Beispiele für die Disposition für Krankheit modulierende CNVs sind Glomerulonephritis [15, 16], Morbus Crohn [17], Autismus [18], Schizophrenie [19] und Psoriasis [20].

Strukturveränderungen, die starken Einfluss auf die Entwicklung eines Menschen haben, und zu zahlreichen Krankheitsbildern führen können, sind lange bekannt [21]. CNVs können aber auch für Krankheiten verantwortlich sein, die nicht in Störungen der individuellen Entwicklung begründet sind, sondern später im Leben auftreten und z. B. neurodegenerative Ursachen haben. So führen eine Triplikation des Alpha-Synuklein-*Locus* zu einem Parkinson-Syndrom mit Lewy-Bodies [22] und eine Duplikation des Alpha-Synuklein-*Locus* zu einer autosomal-dominanten Form des Morbus Alzheimer mit cerebraler amyloider Angiopathie [23].

Im Gegensatz zu Mutationen, die einzelne Basen austauschen oder durch Leserasterverschiebungen die codierenden Abschnitte eines Genes und damit das Protein in seiner Struktur verändern können, haben Kopienzahlveränderungen in der Regel keinen direkten Einfluss auf die Proteinstruktur. Der offensichtlichste und

auch zuerst beschriebene Mechanismus, über den Kopienzahlveränderungen Einfluss auf den Phänotypen haben, ist der Gen-Dosis-Effekt. Dabei können der Verlust oder der Zugewinn eines Genes und die daraus resultierende verminderte bzw. vermehrte Expression seines Produktes weitreichende Folgen für einen Organismus haben. Ein Beispiel für ein solches Dosis-sensitives Gen ist PMP22. Eine Studie, die sich mit dem Einfluss von SNPs und CNVs auf die Genexpression befasst hat, konnte eine klare Korrelation zwischen Strukturvarianten und Expressionsleveln zeigen [24].

## 1.4 Entdeckung von Strukturvarianten und „Database of Genomic Variants“

Im Jahre 2004 wurden in zwei unabhängigen Studien [25, 26] Kopienzahlveränderungen in “normalen“, gesunden Individuen entdeckt. Beide Gruppen verwendeten aCGH (Array comparative genomic hybridization), eine Methode, bei der eine fluoreszenz-markierte Proben-DNA und eine Kontroll-DNA auf einem DNA-Mikrochip hybridisiert werden und dann die Differenz der Intensitäten bestimmt wird. Die erste Gruppe fand 221 CNVs in 20, die zweite Gruppe 255 CNVs in 55 Individuen. In 2005 hat ein Projekt, das die Enden einer Fosmid-Bibliothek sequenziert hat, 297 Strukturvarianten gefunden, in denen sich die Sequenz der Proben-DNA eines einzelnen Individuums von der Referenzsequenz des Human Genome Project unterscheidet [27]. Darunter waren auch 56 Inversionen. Weitere Studien folgten und brachten eine nicht geahnte Anzahl von Strukturvarianten zu Tage. In 2006 wurde gezeigt, dass die SNP-Genotypisierung eines Individuums auch Rückschlüsse auf seine Strukturvarianten zulässt [28-30]. Ende 2006 wurden beide Methoden, aCGH und SNP-Genotypisierung, in einer herausragenden Arbeit [31] kombiniert. Dabei wurden 270 Individuen, die Teil des HapMap-Projekts (s. u.) sind, genomweit auf CNVs hin untersucht, und es wurden 1447 kopienzahl-variable Regionen gefunden. Diese umschreiben Regionen im menschlichen Genom, die überlappende CNVs beherbergen können, und werden im weiteren CNVR (Copy Number Variable Region) genannt. Insgesamt wurden 360 Mb bzw. ca. 12 % des Genoms beschrieben, die variabel in ihrer Kopienzahl sein können. In 2007 konnte durch die Kombination von Endpaar-Konstrukten und Hochdurchsatz-Sequenziermethoden (Paired-end Mapping) [32] die Sensitivität und Auflösung zum Auffinden von

Strukturvarianten drastisch erhöht werden. Eine Studie für zwei Individuen fand über 1000 Strukturvarianten (472 bzw. 825), in denen sich die beiden Individuen von der Referenzsequenz unterschieden.

Je mehr Strukturvarianten gefunden wurden, desto größer wurde der Bedarf nach einer einheitlichen Sammlung aller bekannten Varianten. Die wohl umfassendste Datenbank ihrer Art ist die „Database of Genomic Variants“ (DGV) [26] des „Centre for Applied Genomics“ der Universität von Toronto. Hier werden alle Ergebnisse der neuesten Studien eingetragen. Am 10. November 2008 umfasste die DGV basierend auf 28 zitierten Artikeln insgesamt 31615 Einträge, die sich in 19792 CNVs, 487 Inversionen und 11336 InDels (InDels sind Insertionen oder Deletionen, die kleiner sind als 1 Kb) gliedern. Die CNVs werden zu 6225 *Loci* zusammengefasst.

Abbildung 2 gibt eine Übersicht der Größenverteilung der in der DGV annotierten CNVs. Der relative hohe Wert in der Gruppe von 100 bis 200 Kb spiegelt eine Verzerrung wider, die auf Daten beruht, die mittels BAC-Mikrochips gewonnen wurden. „Bacterial-Artificial-Chromosomes“ (BACs) sind Vektoren zur Klonierung menschlicher DNA in Bakterien. Diese tragen einen großen Teil zu den Daten in der DGV bei. Da diese Studien die Größe der BAC-Klone als die Grenzen der gefundenen CNVs nennen, obwohl eine kleinere Variante gefunden wurde, gibt es eine künstliche Verschiebung hin in den Bereich um 150 Kb.

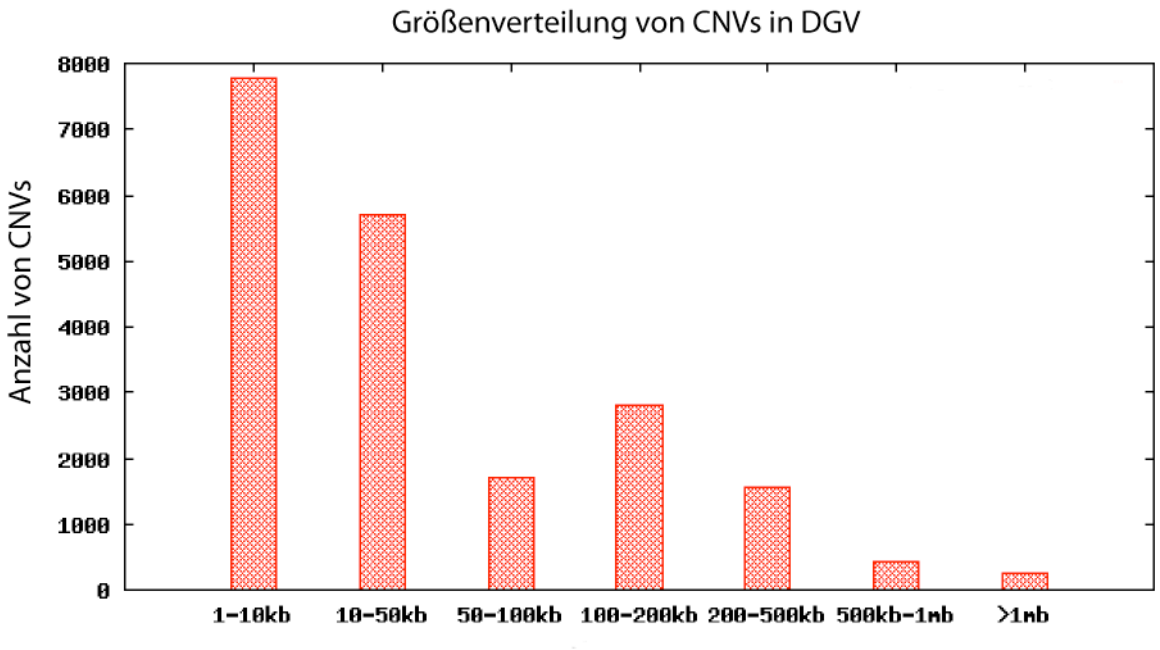



**Abbildung 2: Größenverteilung von CNVs in der DGV**
Die Grafik zeigt die Zuordnung von CNVs zu bestimmten Größenkategorien. Beachte ein relatives Maximum bei 100-200 Kb.

## 1.5 ENCODE und HapMap

Im Folgenden werden zwei Forschungsprojekte beschrieben, die für meine Studie bedeutsam sind. Es handelt sich dabei um „ENCODE“ (**ENC**yclopedia **O**f **D**NA **E**lements) und „HapMap“ (International Haplotype Map Project).

ENCODE [33] ist ein öffentliches Forschungskonsortium, dessen Ziel es ist, alle funktionellen Elemente im menschlichen Genom zu finden. Das Projekt wurde im September 2003 vom “US National Human Genome Research Institute (NHGRI)” gegründet und umfasst mehrere kollaborierende Arbeitsgruppen in den USA und Großbritannien. Die Daten, die im Laufe des Projektes zusammengetragen werden, werden zu festgesetzten Terminen während des Projektes in öffentlichen Datenbanken zugänglich gemacht. Damit soll ermöglicht werden, dass Forscher in der ganzen Welt Zugang zu einheitlichen, vergleichbaren Daten haben, die sonst nicht erhältlich wären.

Da eine genomweite Analyse aller funktionellen Elemente aufgrund technischer und finanzieller Limitierung nicht ohne weiteres möglich ist, konzentriert sich dieses Projekt auf etwa 1 Prozent des menschlichen Genoms. Diese etwa 30 Megabasen (Mb) sind über 44 Regionen genomweit verteilt. Die eine Hälfte wurde zufällig

ausgewählt, die andere umfasst manuell ausgewählte Regionen von besonderem Interesse. Diese ENCODE-Regionen wurden im Rahmen des ENCODE-Projekts nach folgenden Kriterien ausgewählt: 50 % wurden manuell ausgesucht, entweder weil sie gut untersuchte Gene oder andere bekannte Sequenzelemente enthielten oder weil für diese Regionen bereits eine substantielle Menge an Sequenzdaten vorlag, die für vergleichende Untersuchungen genutzt werden konnten. Die andere Hälfte wurde zufällig ausgesucht, basierend auf einer geschichteten Zufallsauswahl, die Gen-Dichte und nicht-exonische Konservierung berücksichtigte. Die 14 manuell ausgewählten Regionen haben jeweils eine Größe zwischen 500 Kb und 2 Mb, die 30 zufällig ausgewählten sind jeweils 500 Kb groß.
Diese verschiedenen Regionen werden in verschiedenen Zelllinien u. a. auf Genexpressionsmuster, Transkriptionsfaktorbindungsverhalten und Histonmodifikationen hin untersucht, um die der Genregulierung zugrunde liegenden Mechanismen besser zu verstehen.

HapMap [3-5] ist ebenfalls ein internationales Konsortium, dessen Ziel es ist, genetische Ähnlichkeiten zwischen Menschen zu entdecken und katalogisieren. Die gewonnenen Informationen werden im Verlaufe des Projektes der Öffentlichkeit zugänglich gemacht. Sie sollen Forschern in der ganzen Welt als Ressource dienen, Gene zu finden, die Einfluss auf Gesundheit, Krankheit und Therapiemethoden haben.
Im Rahmen des Projektes werden die genetischen Sequenzen verschiedener Individuen verglichen und die chromosomalen Regionen identifiziert, die genetische Varianten teilen. Bei diesen Varianten handelt es sich um SNPs, die zu einem Haplotypen zusammengefügt werden können und Rückschlüsse auf die Lage von krankheitsverursachenden Genen zulassen.
In der ersten Phase des HapMap-Projekts wurden Individuen aus vier verschiedenen Ethnien untersucht. Dabei handelt es sich um Yoruber aus Nigeria, Chinesen, Japaner und Kaukasier aus Utah (USA). Mehrere hundert Individuen haben Blut gespendet, das genutzt wurde, um lymphoblastoide Zelllinien zu etablieren. Diese können in Zellkultur gezüchtet werden, um die DNA des Spenders zu gewinnen, und sind aufgrund der umfassenden Informationen zu ihren Genotypen eine wertvolle Grundlage für viele Genomikstudien.

# 2 Hypothese

Ein wichtiger Schritt für die vollständige Erforschung des Phänomens von Kopienzahlveränderungen und ihrer Bedeutung für phänotypische Variationen und Krankheiten ist eine vollständige Kartierung aller CNVs im menschlichen Genom. Mit diesem Katalog kann dann die Gesamtheit aller CNVs eines einzelnen Individuums verglichen werden. Dadurch wird es möglich, Rückschlüsse auf krankheitsverursachende Variationen zu ziehen, die zuvor nicht untersucht werden konnten. Dafür bedarf es einer Methode, die mit hoher Genauigkeit auch kleinste und bisher nicht beschriebene CNVs feststellen kann. Die Methoden, um CNVs am effektivsten aufzufinden, sind konventionelle aCGH und SNP-Genotype *Arrays*. Die erste Methode weist bisher genomweit meist eine recht niedrige Auflösung auf, und die SNP-*Arrays* haben neben einer mittleren Auflösung von etwa 50 Kb eine Verzerrung zugunsten bestimmter Regionen im Genom. Viele komplexe Regionen sind in SNP-Assays meist unterrepräsentiert, da SNPs in komplexen Regionen schwieriger zu bestimmen sind und deshalb seltener die Qualitätskontrollen für einen *Array* bestehen. Aber gerade diese Regionen sind anfällig für Strukturvarianten und CNVs. Beide Methoden haben also Limitierungen in ihrer Möglichkeit, kleine und neue CNVs zu finden und übersehen daher potentiell wichtige Informationen.

Hier stelle ich eine detaillierte Untersuchung aller 44 ENCODE-Regionen vor, welche etwa 1 Prozent des menschlichen Genoms entsprechen. Ich verwende dafür eine hochauflösende Form der aCGH (High-Resolution CGH (HR-CGH)) [34, 35], bei der die zu untersuchende genomische Sequenz durch einen überlappenden Pfad von Oligomeren auf einem Mikrochip repräsentiert wird. Dadurch sollte in diesem Teil des Genoms eine bisher nicht dagewesene Auflösung für die Detektion von CNVs erreicht werden können. Davon verspreche ich mir, zum einen eine größere Anzahl v. a. kleinerer und vorher nicht beschriebener CNVs zu finden und zum anderen deren Ausdehnung genauer als bisher bestimmen zu können.

Anschließend werde ich meine Ergebnisse mit denen anderer Studien vergleichen und evaluieren, ob diese hochauflösende Methode Vorteile beim Auffinden von CNVs hat, wie z. B. eine erwartete höhere Präzision in der Vorhersage der exakten Ausdehnung des CNV und eine niedrige Rate an falsch-negativen Ergebnissen.

Bisher wird davon ausgegangen, dass etwa 12% des menschlichen Genoms variabel in Kopienzahl sind. Da die in meiner Studie angewandte Methode eine etwa 50-mal höhere Auflösung als die meisten der zuvor benutzten Methoden aufweist, erwarte ich eine deutliche Reduktion dieser Schätzungen.
Um diese Methode besser auf ihre Fähigkeit, neue CNVs zu finden, zu testen, füge ich meinem Probensatz neben Individuen aus Populationen, die bisher mehrfach in anderen Studien untersucht wurden, auch Individuen anderer, bisher weniger gut untersuchter Populationen hinzu. Damit werden voraussichtlich die genetische Variation und die Möglichkeit, neue CNVs zu finden, erhöht.

# 3 Material und Methoden

## 3.1 Probensatz

In dieser Studie werden insgesamt 36 verschiedene phänotypisch unauffällige Individuen auf Kopienzahlveränderungen hin untersucht. 16 dieser Proben wurden zuvor in den zwei Phasen des HapMap-Projekts intensiv untersucht und ihr detaillierter SNP-Genotyp ist bekannt. Es handelt sich dabei um fünf Kaukasier, fünf Asiaten (Chinesen und Japaner) und fünf Yoruber aus Nigeria. Einige dieser HapMap-Proben wurden in den letzten Jahren in verschiedenen weiteren Studien intensiv auf CNVs hin untersucht und dienen deshalb oft als Referenzen für verschiedene Methoden und ihre Evaluierung. Abbildung 3 und Tabelle 1 zeigen die unterschiedlichen Regionen der Welt, aus denen sich mein Probensatz rekrutiert.

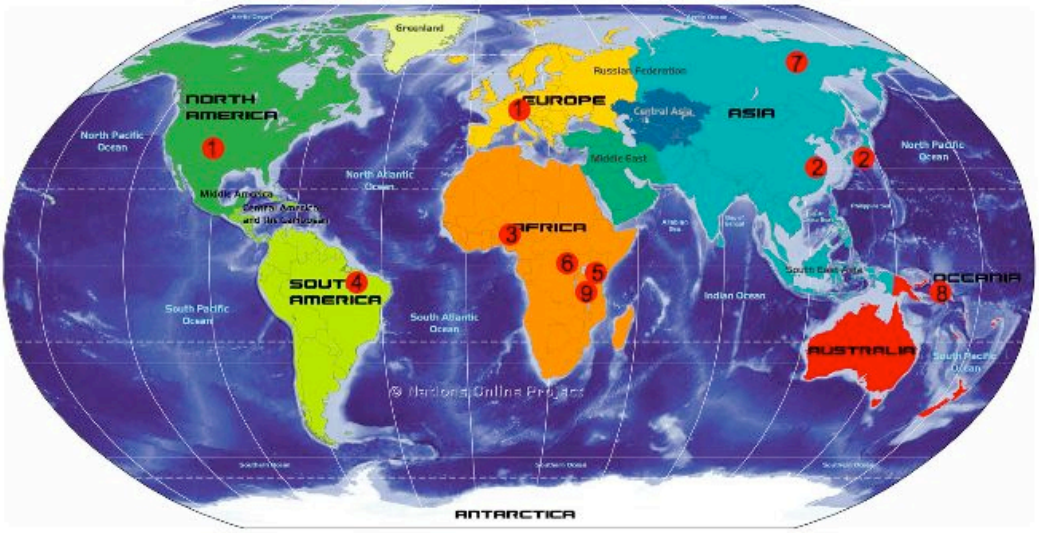



**Abbildung 3: Herkunft der Probensätze**
Geografische Zuordnung der in dieser Arbeit verwendeten Probensätze. Insgesamt fanden Probensätze neun unterschiedlicher Ethnien (1-9) Verwendung.

| Nummer auf Weltkarte | Ethnie | Region | Anzahl an Individuen |
|---|---|---|---|
| 1 | Kaukasier (CEPH) | Nord-Amerika, Europa | 7 |
| 2 | Asiaten (JCH) | Japan, China | 5 |
| 3 | Yoruber (YRI) | Zentral-Afrika, Nigeria | 5 |
| 4 | Surui | Nord-Brasilien | 2 |
| 5 | Masai | Kenia, Tansania | 3 |
| 6 | Mbuti | Ost-Kongo | 3 |
| 7 | Yakut | Zentral-Russland | 3 |
| 8 | Nasioi | Papa Neu-Guinea | 3 |
| 9 | Sandawe | Dodoma Region in Zentral-Tansania | 5 |

**Tabelle 1: Detaillierte Informationen über die Herkunft der Probensätze**
Insgesamt wurden Probensätze von 36 Individuen, die zu neun unterschiedlichen Ethnien gehören, verwendet.

Eine weitere Probe habe ich vom Corriel Cell Repositorium [36] bezogen. Es handelt sich dabei um ein Individuum (15510), das als erstes mit einem Paired-end Mapping-Ansatz [27] untersucht wurde. Es ist außerdem zusammen mit einem weiteren Individuum (18505) in meiner Studie mit einer weiteren Paired-end Mapping-Methode untersucht worden. Dabei wurden ungefähr 450 bzw. 800 Strukturvarianten in den beiden Individuen im Vergleich zur Referenzsequenz gefunden [32]. Die restlichen 19 Proben repräsentieren gesunde, bisher wenig untersuchte Individuen aus verschiedenen, teilweise isolierten Populationen (Surui, Nasioi, Masai, Mbuti, Sandawe, Yakut). Mit diesen sollte die genetische Diversität der Studie erhöht werden.

## 3.2 aCGH und Microarray-Aufbau

Prinzip:

Das den *Microarrays* zugrunde liegende Prinzip ist die Hybridisierung zweier komplementärer DNA-Stränge. Der eine Strang ist eine bekannte Sonde, die – im Gegensatz zum *southern blot* – auf einem Objektträger gebunden ist. Der andere Strang ist Teil der zu untersuchenden DNA. Dieses *Target* kann sich spezifisch an die Sonde lagern, wobei Wasserstoffbrücken zwischen komplementären Basenpaaren ausgebildet werden. Je größer die Anzahl komplementärer Basen in zwei einzelsträngigen DNA-Molekülen ist, desto stärker ist die Bindung zwischen den beiden Strängen. Nach der Hybridisierung wird der *Microarray* gewaschen, um unspezifische, weniger starke Bindungen zu entfernen. Wenn die zu untersuchende DNA vorher mit z. B. einem fluoreszierenden Farbstoff markiert worden ist, kann dann mit einem *Scanner* die Signalstärke für jede Bindung zwischen Sonde und gebundenem *Target* bestimmt werden. Da Position und Eigenschaften der verschiedenen Sonden auf dem Objektträger bekannt sind, kann man nun Rückschlüsse auf die untersuchte DNA ziehen.

Im Falle von Nimblegen *Arrays* werden zwei verschiedene Sätze von DNA mit zwei verschiedenen Farbstoffen markiert (s. Abbildung 4). Der erste Satz ist die zu untersuchende DNA, der zweite eine Kontroll-DNA, deren Eigenschaften entweder bekannt sind oder durch das „*pooling*“ mehrerer Kontrollindividuen einem Durchschnitt entsprechen sollen. Während der Hybridisierung konkurrieren beide DNA-Proben um die komplementären Sonden auf dem Objektträger. Anschließend wird mit einem Zwei-Kanal-*Scanner* die Differenz zwischen den Signalen beider Fluorochrome bestimmt, um Unterschiede in der DNA-Struktur zu messen. Somit lassen sich relative Unterschiede in der Kopienzahl in genomischer DNA feststellen.

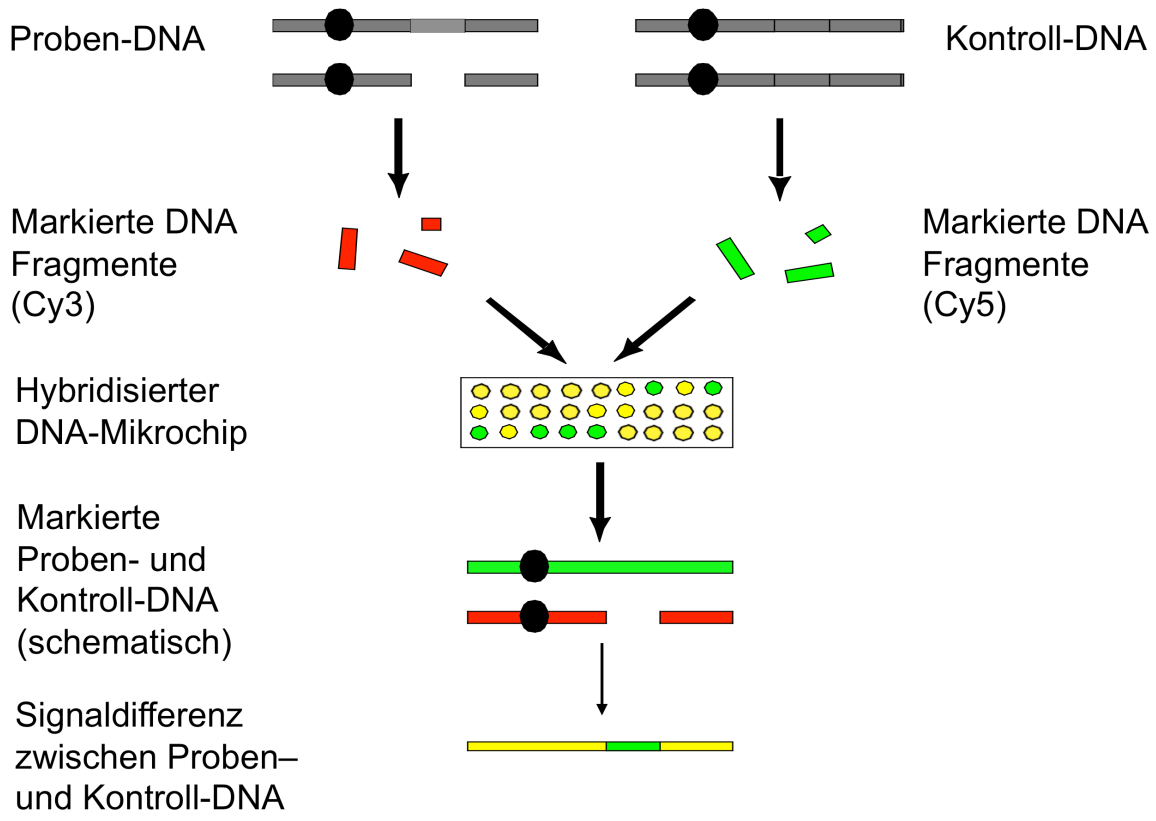



**Abbildung 4: Prinzip der Microarray Hybridisierung**
Schematische Darstellung: zwei unterschiedlich markierte Proben werden auf einem Mikrochip hybridisiert; anschließend werden die Signalintensitäten bestimmt.

Aufbau von Nimblegen *Arrays*:

Die in dieser Studie benutzten *Arrays* wurden von NimbleGen-Systems (jetzt Roche-NimbleGen, Madison, WI) [37] erworben. Es handelt sich um HR-aCGH *Arrays*, die mittels maskenloser Synthese (MAS, “maskless synthesis”) mit Photolithographie [38] hergestellt wurden. Jeder *Array* hat 385000 Oligomer-Sonden, die so ausgewählt wurden, dass sie den nicht-repetitiven Teil der 44 ENCODE Regionen, die 1 Prozent des menschlichen Genoms ausmachen, repräsentieren. Die einzelnen Oligomer-Sonden sind 50-mere, die alle 38 bp entlang der genomischen Sequenz beginnen und daher eine überlappende Abdeckung ergeben. Die repetitiven Anteile der genomischen Sequenz wurden zuvor mit dem RepeatMasker-Algorithmus [39] identifiziert und ausgeschlossen, um mögliche Kreuzhybridisierungen zu minimieren.

## 3.3 Probenvorbereitung und Hybridisierung

Genomische DNA wurde nach folgendem Protokoll [34, 35] vorbereitet: genomische DNA wurde per Sonikation geschert. DNA (2 µg pro Kanal auf dem *Array*) wurde nach Denaturierung mit Cy3- oder Cy5-konjugierten Nonameren mit randomisierter Sequenz (TriLink Biotechnologies, San Diego), 100 U (Exo-) Klenow Fragment (NEB, Beverly, MA) und dNTP [6 mM für jedes der vier Nukleotide, in TE Puffer (10 mM Tris, 1 mM EDTA, pH 7,4; Invitrogen)] inkubiert, für 2 Stunden bei 37º C.
Die Reaktionen wurden durch Zugabe von 0,5 M EDTA (pH 8,0) beendet, DNA wurde mit Isopropanol ausgefällt und in Wasser resuspendiert. Die Cy-konjugierten Proben (Cy3) und eine Referenzmischung bestehend aus genomischer DNA von sieben normalen männlichen Probanden (Promega) (Cy5) wurden kombiniert in NimbleGen Hybridisierungspuffer (Roche-NimbleGen). Nach Denaturierung erfolgte die Hybridisierung mit dem *Array*, durchgeführt auf einem MAUI-Hybridisierungssystem (BioMicro Systems, Salt Lake City), für 18 Stunden und 42º C. Die *Arrays* wurden mit NimbleGen-Waschpuffern gewaschen, durch Zentrifugation getrocknet und mit einem GenePix 4000B Scanner (Axon Instruments, Union City, CA) bei 5 µm Auflösung eingelesen.

## 3.4 Analyse der Microarray-Daten

Die Fluoreszenz-Intensitätsdaten habe ich mit der „NIMBLESCAN 2.0 extraction software“ (NimbleGen Systems) von den gescannten Bildern der Oligonukleotid-*Arrays* erhalten. Für jeden Punkt auf dem Chip wurden die log2-ratios zwischen der Cy3-markierten Proben-DNA und der Cy5-markierten Referenz-DNA berechnet.
Diese Daten wurden mit der Nimblescan-Software normalisiert und dem implementierten SegMNT-Algorithmus bewertet. Die ausgegeben Dateien wurden anschließend mit der SignalMap-Software (NimbleGen Systems) visualisiert.

## 3.5 Methoden zur Validierung der gefundenen CNVs

### 3.5.1 Quantitative real-time PCR (qPCR) und relative Quantifizierung

Die qPCR ist z. Zt. die genaueste Methode, um eine gegebene Anzahl von DNA-Molekülen zu bestimmen, und gilt daher als der “Goldstandard”, um Kopienzahlveränderungen quantitativ zu erfassen. Die qPCR beruht auf den Grundlagen der Polymerasekettenreaktion (PCR). Im Reaktionsansatz befindet sich

ein spezieller fluoreszierender Farbstoff (SYBRGreen), der nur angeregt werden kann, wenn er in eine DNA-Helix eingelagert ist. Die anschließend vom SYBRGreen wieder abgegebene Strahlung wird von einer Kamera in der PCR-Maschine gemessen. Die Intensität der abgegebenen Strahlung ist proportional zur Anzahl der vorliegenden DNA-Moleküle im Ansatz. Mit jedem Zyklus der Reaktion verdoppelt sich idealerweise die Zahl der DNA-Moleküle, und somit nimmt auch die Intensität der Strahlung exponentiell zu. Den effektiven Messwert für jede einzelne Reaktion erhält man, indem man einen für alle Reaktionen gleichen Intensitätswert festlegt. Dann bestimmt man für jede Reaktion den Zyklus, mit dem dieser Schwellenwert erreicht wurde. Dieser Wert ist der Ct-Wert einer Reaktion.

Um in einem Individuum die Kopienzahl an einem bestimmten *Locus* im Genom zu bestimmen, habe ich den relativen Unterschied zwischen der jeweiligen Proben-DNA und einer Kontroll-DNA berechnet. Da die jeweilige Kopienzahl in der Kontroll-DNA unbekannt ist, habe ich einen DNA-Pool von sieben verschiedenen Individuen benutzt, um Fehler durch eventuell vorhandene Kopienzahlveränderungen in der Kontroll-DNA zu minimieren. Für jeden zu testenden Ziel-*Locus* (z) wurden ein oder mehrere spezifische Primerpaare entworfen. Um die Unterschiede in DNA-Konzentration zwischen Test- und Kontroll-DNA zu normalisieren, wurde immer dasselbe Primerpaar für einen Referenz-*Locus* (r) verwendet, von dem angenommen wurde, dass er nicht variabel in seiner Kopienzahl ist. Die Normalisierung berücksichtigt die Effizienz der verschiedenen Primerpaare und wurde gemäß folgender Formel vorgenommen [40]:

$$\text{Relative Kopienzahl} = \frac{E_z^{\Delta Ct(\text{Kontrolle}-\text{Proband})z}}{E_r^{\Delta Ct(\text{Kontrolle}-\text{Proband})_r}}$$

Die Effizienz für jedes Primerpaar wurde gemäß folgender Formel bestimmt:

$$\text{Effizienz} = 10^{(-1/\text{Steigung})}$$

Dabei ist „Steigung“ die Steigung einer Standardkurve, die in jedem Lauf anhand einer Verdünnungsreihe berechnet wurde. Diese Verdünnungsreihe umfasste jeweils 6 Verdünnungen im Verhältnis 1:10, mit einer Startkonzentration von 100 ng genomischer DNA.

Alle Experimente wurden mit einem Roche LightCycler 480 mit einem SYBRGreen I MasterMix (Katalognummer 04707516001, Indianapolis, IN) in Triplikaten ausgeführt. Jeder Ansatz auf einer 384-er Platte enthielt ein Endvolumen von 20 µl mit 6 ng genomischer DNA. Die Endkonzentration der spezifischen Primer war 400 nM. Die Reaktionen wurde mit folgendem Protokoll durchgeführt: (1) 5 min. bei 95° C; (2) 45 Zyklen von je 5 sec. bei 95° C, 10 sec. bei 60° C, 30 sec. bei 72° C.

### 3.5.2 Konventionelle Polymerasekettenreaktion (PCR)

Einen Teil meiner Vorhersagen habe ich mit regulärer PCR überprüft. Dabei wurden die Primer so entworfen, dass sie außerhalb der Vorhersage lagen. Im Falle einer Deletion erhält man so ein kürzeres PCR-Fragment als durch *in silico*-PCR vorhergesagt, im Falle einer Duplikation ein längeres. Abbildung 5 illustriert die Bandenmuster, die auftreten können, wenn eine Deletion untersucht wird. Dabei wird davon ausgegangen, dass die zu untersuchende Deletion kleiner als ~ 6 Kb ist, da man sonst im nicht deletierten Zustand keine Bande erhält. Das Muster würde dann nur den deletierten Zustand zeigen (kleinere Bande).

Die Reaktionen wurde mit folgendem Protokoll durchgeführt: (1) 5 min. bei 95° C; (2) 36 Zyklen von je 10 sec. bei 95° C, 30 sec. bei 60° C, 6 min. bei 72° C; (3) 30 min. bei 72° C.

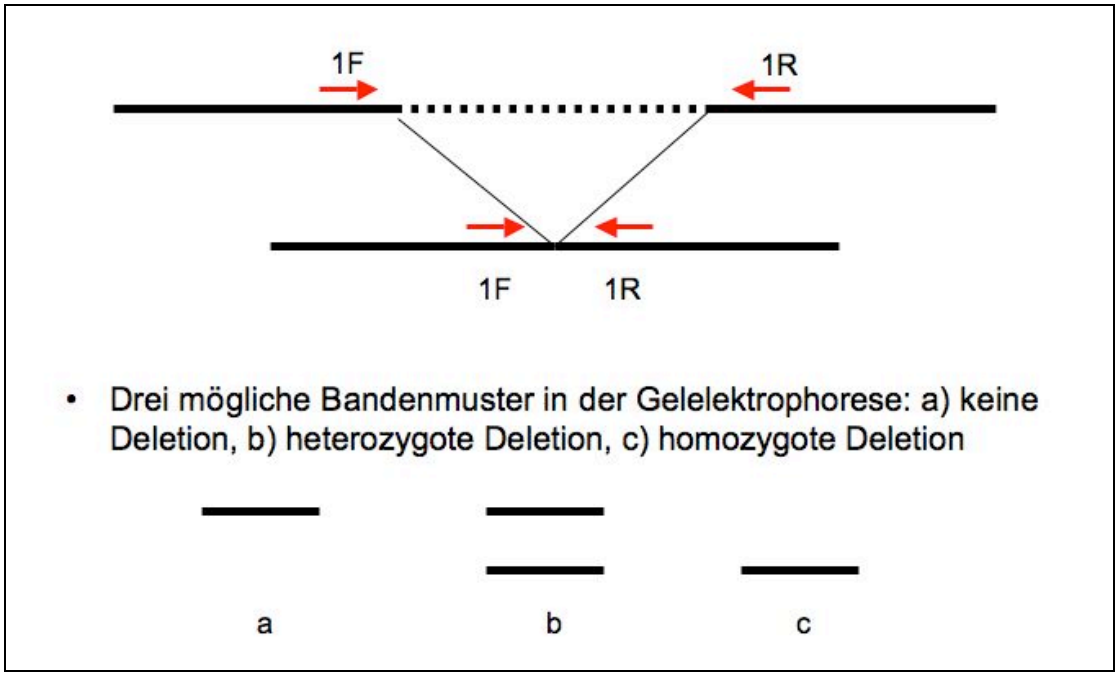



**Abbildung 5: Potentielle Produkte der PCR**
Mögliche Bandenmuster bei der Untersuchung von Deletionen. Je nachdem, ob eine heterozygote (b) oder homozygote Deletion (c) vorliegt, findet man zwei oder nur eine Bande. Liegt keine Deletion vor (a), findet man ebenfalls nur eine Bande, die aber höher läuft als im Falle einer homozygoten Deletion (c).

# *4 Ergebnisse*

## 4.1 aCGH-Ergebnisse

Für meine Untersuchung von CNVs im menschlichen Genom habe ich mich auf die 44 Regionen im menschlichen Genom konzentriert, die für das ENCODE-Projekt ausgewählt wurden. Wie in der Einleitung beschrieben umfassen diese ENCODE-Regionen 30 Mb, welche etwa 1 Prozent des gesamten menschlichen Genoms entsprechen. Die eine Hälfte wurde zufällig ausgewählt, die andere umfasst manuell ausgewählte Regionen von besonderem Interesse. Die ausgewählten Regionen haben jeweils eine Größe zwischen 500 Kb und 2 Mb. Um Kopienzahlveränderungen für diese Region in meinen 36 Test-Individuen zu finden, habe ich einen Nimblegen-Mikrochip benutzt, der alle 44 ENCODE-Regionen umfasst. In jedem Hybridisierungsexperiment wurden die zu untersuchende fluoreszenz-markierte DNA und eine ebenfalls fluoreszenz-markierte Kontroll-DNA auf einen Mikrochip hybridisiert. Die Kontroll-DNA in jedem Experiment war ein DNA-Pool von sieben verschiedenen Individuen.

Insgesamt habe ich 31 verschiedene *Loci* gefunden, die eine CNV in mindestens einem von 36 untersuchten Individuen aufwiesen. Auf diese 31 *Loci* verteilen sich 231 CNVs, davon zeigen 181 eine Abnahme in Kopienzahl (*loss*) und 50 eine Zunahme (*gain*), also eine Verhältnis von fast 4:1. Durchschnittlich zeigt ein Individuum 6,42 CNVs, wobei das Spektrum von 0 bis zu 16 CNVs in einem Individuum reicht. Die gefundenen CNV-*Loci* sind auf 12 der 44 ENCODE-Regionen verteilt. Die übrigen 32 Regionen zeigen keine Variation in meinem Probenset (s. Tabelle 2).

| | |
|---|---|
| ENCODE-Regionen | 44 |
| Davon zeigen CNV | 12 |
| Davon zeigen keine CNV | 32 |

**Tabelle 2: CNV-*Loci* in ENCODE-Regionen**
Die Tabelle zeigt die Anzahl der ENCODE-Regionen, die CNVs aufweisen. Insgesamt zeigen 12 der 44 untersuchten ENCODE-Regionen CNVs.

Die Mehrzahl der betroffenen Regionen zeigen nur einen einzigen *Locus*, der von CNVs betroffen sein kann, während sechs Regionen zwischen zwei und acht verschiedene *Loci* aufweisen können (s. Tabelle 3).

Interessanterweise sind nur vier der zufällig ausgewählten ENCODE-Regionen von CNVs betroffen. Sieben CNVs liegen in diesen vier Regionen, die übrigen 24 CNVs finden sich in acht manuell ausgewählten Regionen. Da die verschiedenen ENCODE-Regionen unterschiedliche Größen haben, ergibt sich folgendes Verteilungsmuster: sieben CNVs in vier zufällig gewählten Regionen mit etwa 2 Mb Größe und 24 CNVs in acht manuell gewählten Regionen mit einer Größe von etwa 8,8 Mb.

| Region | unterschiedliche CNVs gefunden | Basen betroffen |
|---|---|---|
| ENm004 | 1 | 76 |
| ENm005 | 1 | 570 |
| ENm007 | 3 | 51300 |
| ENm008 | 4 | 3386 |
| ENm009 | 4 | 20910 |
| ENm011 | 8 | 102596 |
| ENm013 | 1 | 1748 |
| ENm014 | 2 | 5472 |
| ENr132 | 4 | 18768 |
| ENr231 | 1 | 2400 |
| ENr313 | 1 | 6498 |
| ENr332 | 1 | 1786 |
| **total** | **31** | **205746** |

**Tabelle 3: Anzahl verschiedener CNVs in ENCODE-Regionen**
Die Tabelle listet alle betroffenen Regionen (ENm = manuell ausgewählt, ENr = zufällig ausgewählt), die Anzahl unterschiedlicher CNV-*Loci* und die Anzahl der Basenpaare, die diese *Loci* umfassen, auf.

Insgesamt finde ich ca. 200 Kb, die von CNVs betroffen sein können, in 30 Mb oder 1 Prozent des menschlichen Genoms. Auf das gesamte menschliche Genom

hochgerechnet sind das etwa 20 Mb oder etwa 0,7 Prozent, die von CNVs betroffen sein können. Die Häufigkeit, mit der ein CNV-*Locus* eine CNV aufweist, ist zwischen den unterschiedlichen *Loci* sehr verschieden. Manche *Loci* zeigen nur eine einzige CNV in allen 36 Studienindividuen, bei anderen ist eine Kopienzahlveränderung die häufigere Variante (s. Tabelle 4). So ist z. B. *Locus* ID_11 nur in einem Individuum kopienzahl-variabel, *Locus* ID_19 hingegen in 21 von 36 Individuen.

Tabelle 5 zeigt die Ergebnisse aus Tabelle 4 nach ethnischer Abstammung geordnet. Mehrere Anhäufungen von gleichen CNVs innerhalb einer Ethnie lassen sich ausmachen. Auffällig sind v. a. *Loci* 14-22 in der Gruppe der Yoruber, aber auch in anderen Ethnien sind mehrere Ballungen von CNVs zu finden. Die Gruppe der Yoruber zeigt die meisten CNVs. Es sind im Durchschnitt 14 pro Individuum. Die Gruppe der Japaner und Chinesen zeigt 4,6 CNVs pro Individuum. Überraschend ist, dass die Gruppe der Kaukasier recht wenige CNVs aufweist: 2,14 pro Individuum.

### 4.1.1 CNV-*Loci* in den ENCODE-Regionen

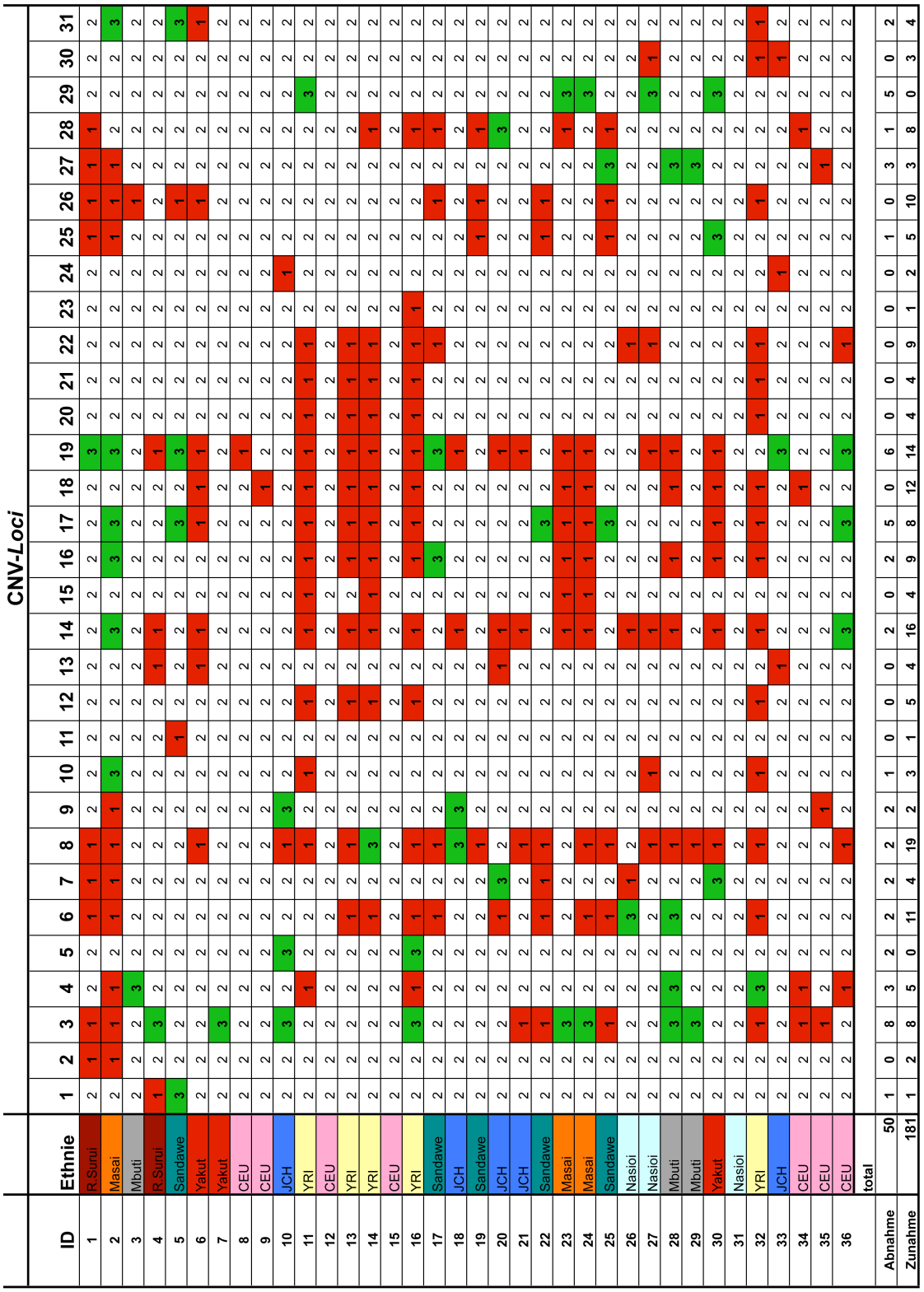



**Tabelle 4: Die 31 CNV-*Loci* in den 36 Individuen**
Für alle 31 CNV-*Loci* sind die Kopienzahlen in den 36 verschiedenen Individuen angegeben. „2“ in weißer Box zeigt eine normale Kopienzahl im Verhältnis zur Kontrolle an, „1“ in roter Box eine Abnahme und eine „3“ in grüner Box eine Zunahme in Kopienzahl.

### 4.1.2 CNV-Gruppierungen geordnet nach Abstammung

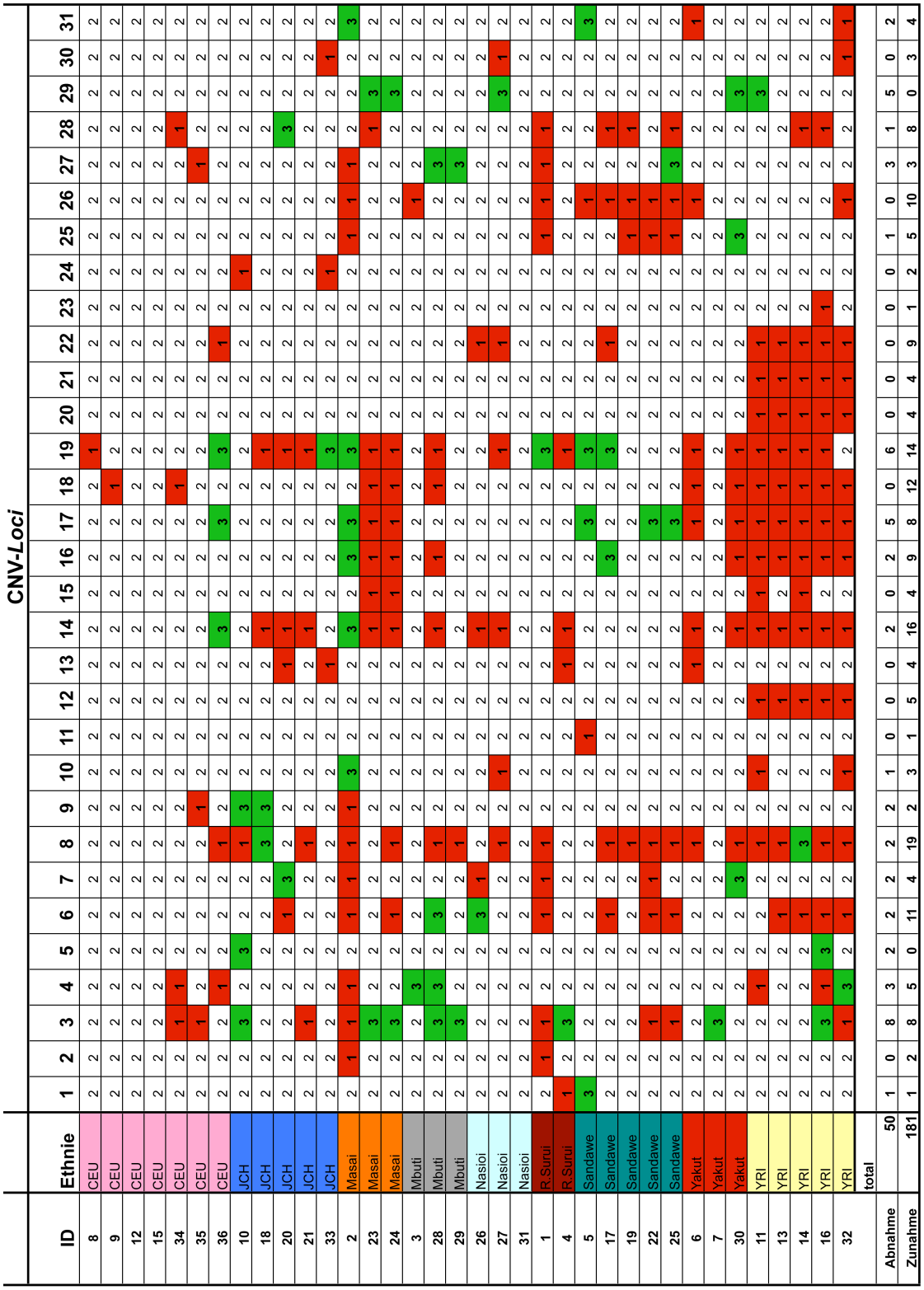



**Tabelle 5: Die 31 CNV-*Loci* nach Abstammung geordnet**
Alle 31 CNV-*Loci* in den 36 verschiedenen Individuen nach ethnischer Abstammung geordnet. Beachte die Anhäufung bestimmter CNVs in verschieden ethnischen Gruppen.

### 4.1.3 Größenverteilung von CNVs

Alle 31 CNV-*Loci* zusammengenommen umfassen ca. 200 Kb genomischer Sequenz. Daraus ergibt sich eine durchschnittliche Größe von etwa 6,5 Kb pro CNV. Das größte CNV umfasst ca. 31 Kb, das kleinste 76 bp. Abbildung 6 gibt eine Übersicht über die Größenverteilung aller CNVs. Die Mehrheit aller Varianten findet sich wie auch in anderen neueren Studien [32] im Spektrum < 10 Kb; fast die Hälfte davon ist sogar kleiner als 1 Kb. Nur fünf der 31 CNVs sind größer als 10 Kb. Größere CNVs sind seltener, sind aber trotzdem für mehr als die Hälfte aller betroffenen Basenpaare verantwortlich.

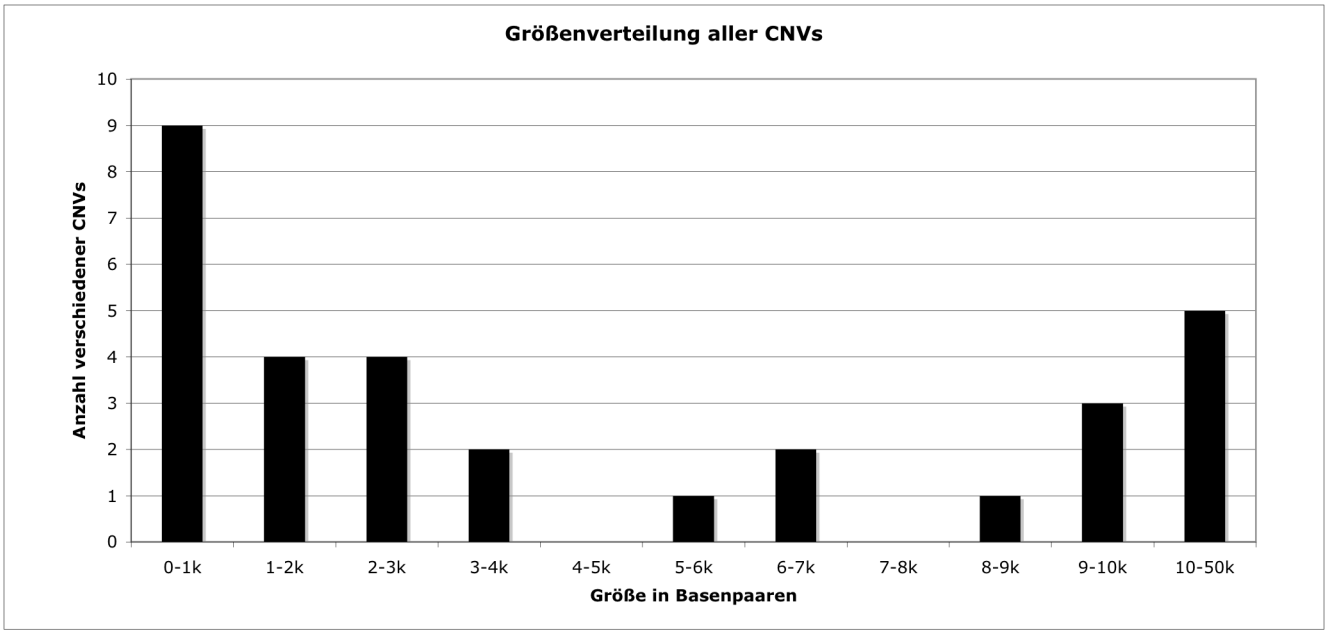



**Abbildung 6: Größenverteilung aller CNVs in den 31 verschiedenen *Loci***
Die Mehrheit der CNVs ist kleiner 4 Kb, davon neun CNVs < 1 Kb. Lediglich fünf CNVs sind größer als 10 Kb.

Die durchschnittliche Größe eines einzelnen CNVs beträgt 6530 bp, der Median liegt bei 2774 bp.

### 4.1.4 Betroffene Gene

Die Mehrheit der hier beschriebenen CNVs (23 von 31) liegt innerhalb von Genen und zwar meist in Introns. Aber fünf dieser 23 CNVs betreffen insgesamt 28 Exons in sieben verschiedenen Genen. Die Tabellen 6 und 7 listen alle CNVs, die innerhalb eines Gens liegen können, auf bzw. die Anzahl der betroffenen Exons und

die Zugehörigkeit zu Genfamilien. Es handelt sich dabei v. a. um Olfaktorische Rezeptorgene oder Gene, die der Immunmodulation dienen. Dies sind Gruppen von Genen, die generell eine höhere Wahrscheinlichkeit eines CNVs aufweisen [12].

| Nr | Region | Name | Exon | Intron |
|---|---|---|---|---|
| 1 | ENm004 | SYN3 | nein | ja |
| 2 | ENm005 | SYN3J1 | nein | ja |
| 3 | ENm007 | CACNG7 | nein | ja |
| 4 | ENm007 | LILRA6 | 6 | ja |
| 5 | ENm007 | LILRB2,LILRA3 | 17 | ja |
| 6 | ENm008 | MPG | nein | ja |
| 9 | ENm008 | RAB11FIP3 | nein | ja |
| 10 | ENm009 | OR51A4,OR51A2 | 2 _fusion | nein |
| 11 | ENm009 | OR51L1 | 1 | nein |
| 12 | ENm009 | HBG2,HBE1,OR51B5 | nein | ja |
| 13 | ENm009 | HBG2,HBE1,OR51B5 | nein | ja |
| 14 | ENm011 | HCCA2, CR626060 | nein | ja |
| 16 | ENm011 | LSP1 | nein | ja |
| 17 | ENm011 | LSP1 | 2 | ja |
| 20 | ENm011 | IGF2,IGF2AS | nein | ja |
| 22 | ENm013 | STEAP2 | nein | ja |
| 23 | ENm014 | GRM8 | nein | ja |
| 25 | ENr132 | ATP11A | nein | ja |
| 26 | ENr132 | ATP11A,KIAA1021 | nein | ja |
| 27 | ENr132 | ATP11A,KIAA1021 | nein | ja |
| 28 | ENr132 | MCF2L | nein | ja |
| 29 | ENr231 | PI4KB | nein | ja |
| 31 | ENr332 | NRXN2 | nein | ja |

**Tabelle 6: Gene mit von CNV betroffenen Exons**
Die Tabelle zeigt alle CNVs, die in einem Gen liegen, die Region, in der sie liegen, den Namen des Gens und ob Exons oder Introns betroffen sind. "2_fusion" steht für ein Fusionsprodukt aus zwei Exons.

| Genname | Beschreibung | Anzahl betroffener Exone |
|---|---|---|
| OR51L1 | olfactory receptor family 51 subfamily L | 1 |
| LSP1 | lymphocyte-specific protein 1 isoform 2 | 2 |
| LILRA6 | leukocyte immunoglobulin-like receptor | 6 |
| LILRB2 | leukocyte immunoglobulin-like receptor | 13 |
| LILRA3 | leukocyte immunoglobulin-like receptor | 4 |
| OR51A4 | olfactory receptor family 51 subfamily A | 1 |
| OR51A2 | olfactory receptor family 51 subfamily A | 1 |

**Tabelle 7: Gene mit von CNV betroffenen Exons**
Die Tabelle zeigt alle Gene mit mindestens einem von CNV-betroffenen Exon, deren Zuordnung zu Genfamilien sowie die Anzahl der betroffenen Exons.

Besonders bemerkenswert ist ein CNV (ID_10), dessen beiden Enden jeweils in einem Olfaktorischen Rezeptorgen zu liegen kommen und damit zu einem Fusionsgen führen, das für eine veränderte Proteinstruktur codiert [32]. Abbildung 7 zeigt die etwa 9700 bp große Deletion und die betroffenen Exons, die innerhalb zweier Segment-Duplikationen liegen.

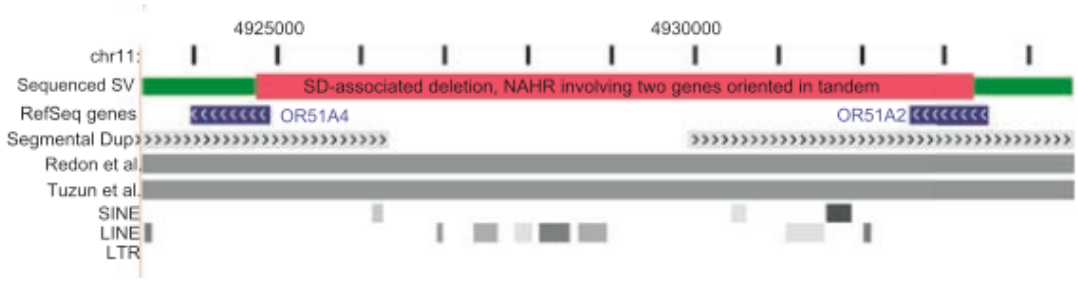



**Abbildung 7: CNV führt zu Fusionsgen**
Der rote Balken markiert ein CNV zwischen zwei olfaktorischen Rezeptrogenen. Die beiden grauen Balken zeigen Annotationen aus zwei anderen Studien für die gleiche Region, mit jedoch geringerer Auflösung.

### 4.1.5 Hypervariable CNVs – Simple Tandem Repeats (STR)

Elf der 31 CNV-*Loci* zeigen einige Besonderheiten: ein besonders deutliches Signal auf dem *Microarray* bei (meist) geringer Größe und hypervariabler Ausprägung der Allelzustände. Abbildung 8 zeigt einen aCGH-Plot für einen hypervariablen *Locus* in fünf verschiedenen Individuen. Drei der Individuen zeigen eine Zunahme in der Kopienzahl (28, 25 und 13), zwei Individuen ein Abnahme in der Kopienzahl (1 und 2). Die Intensität des Signals variert dabei sehr stark.

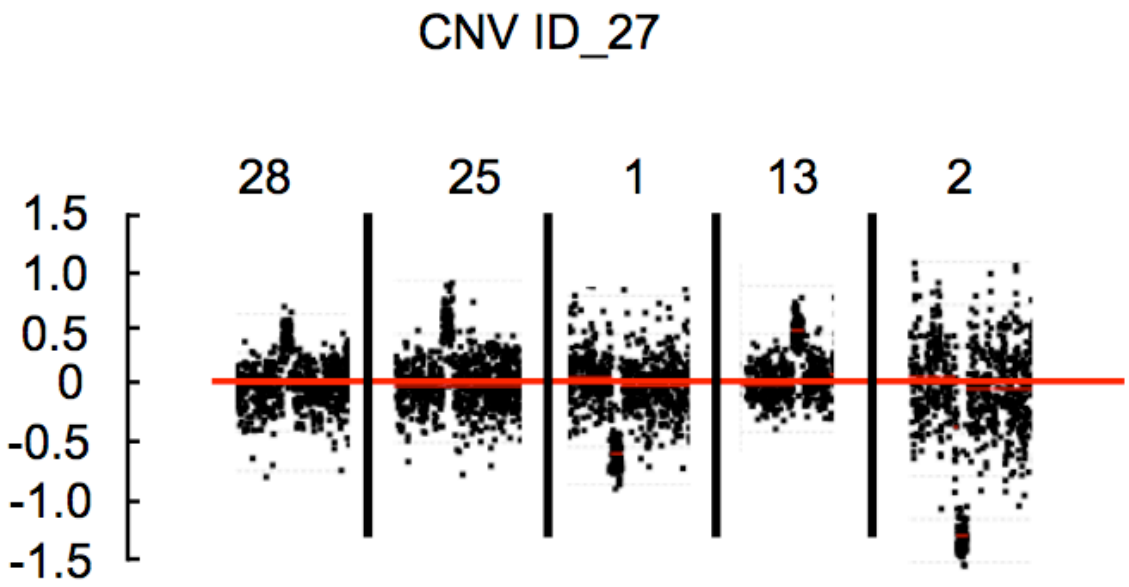



**Abbildung 8: STR-CNV-*Locus* in fünf verschiedenen Individuen**
Signalintensitäten sind auf einer log2-Skala angegeben.

Ich habe für drei dieser *Loci* Primerpaare außerhalb der vermuteten Bruchpunkte entworfen, um anschließend die resultierenden PCR-Amplifikate sequenzieren zu können. Abbildung 9 zeigt die Gelelektrophorese für *Locus* ID_8. Für das Primerpaar war auf dem Referenzgenom basierend durch *in silico* PCR eine Produktgröße von 1748 bp vorhergesagt. Keines der Produkte hat diese Größe. Das kleinste Produkt (32) ist etwa 400 bp und das größte ist etwa 1 Kb groß. Die Mehrzahl der PCR-Produkte hat eine Größe zwischen 500 und 700 bp. Die meisten Bahnen zeigen nur eine Bande (z. B. 3 und 14), was einer Deletion auf beiden Chromosomen entspricht. Wenige Bahnen haben zwei Banden (z. B. 11 und 19) und manche Bahnen zeigen mehrere Produkte (z. B. 16, 17 und 28).

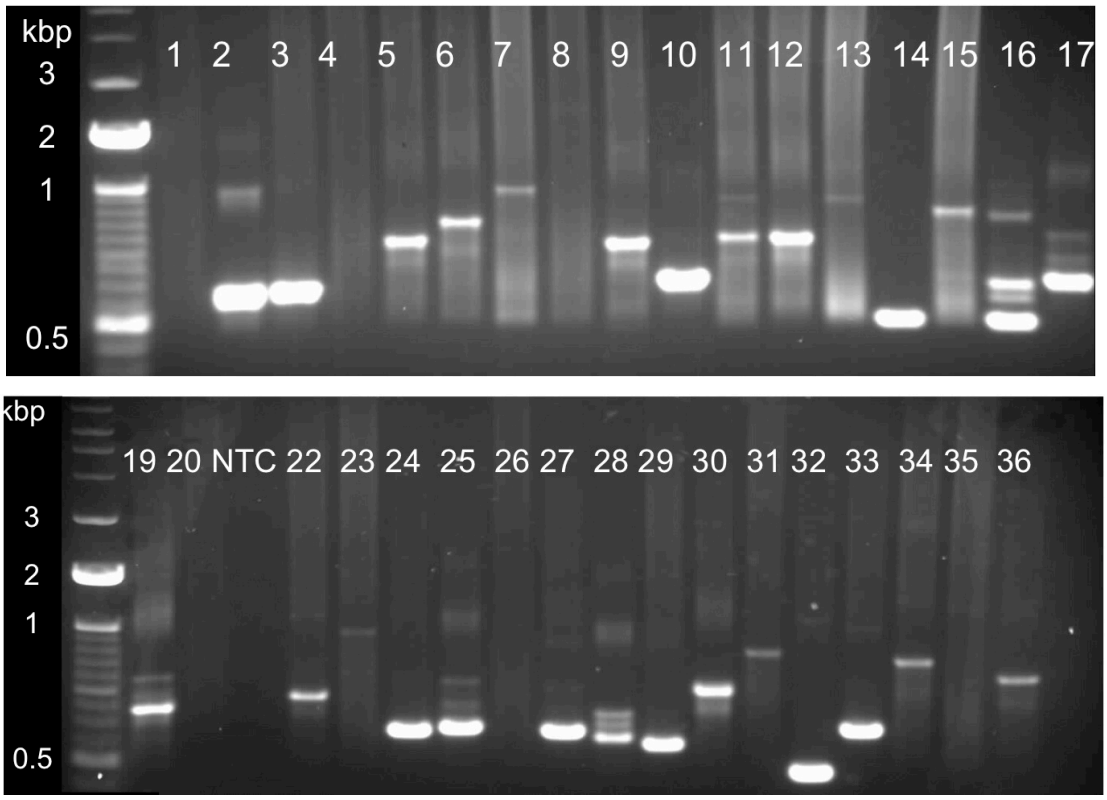



**Abbildung 9: Gelelektrophorese für CNV-*Locus* ID_8**
Die Größen der PCR-Produkte des CNV-*Locus* ID_8 der verschiedenen Individuen variieren stark. Die Negativ-Kontrolle zeigt kein Produkt (NTC).

Ich habe die PCR-Produkte sequenziert und beim Vergleich gefunden, dass alle Fragmente an einem solchen CNV-*Locus* ein gemeinsames zugrunde liegendes Sequenzmerkmal haben: eine Folge sich tandemartig wiederholender, kurzer (~ 17 bp) DNA-Abschnitte. Die Anzahl dieser Wiederholungen bestimmt die Länge des PCR-Produktes und korreliert zur Intensität des Signals auf dem *Microarray*. Die Bruchpunkte der einzelne CNVs sind je nach Anzahl an *Tandem Repeats* sehr variabel.

Abbildung 10 zeigt die Sequenz für zwei Individuen (3 und 7). Beide weisen im Vegleich zur Referenzsequenz eine Deletion auf. Während bei Individuum 3 ein Sequenzblock von ca. 1000 bp fehlt, zeigt Individuum 7 zwei fehlende Blöcke von je etwa 200 bzw. 400 bp Länge.

**Abbildung 10: Vergleich zweier PCR-Amplifikate mit der Referenzsequenz**
Sequenzvergleich zwischen Referenzsequenz (schwarz) und Sequenz des PCR-Produktes von Individuum 3 (links, blau) bzw. Individuum 7 (rechts, blau).

Zwei ENCODE-Regionen weisen besonders häufig diese Art der Variation auf: ENr132 und ENm008. Beide liegen in der Nähe eines Telomers auf Chromosom 13 bzw. Chromosom 16. Abbildung 11 zeigt den aCGH-Plot für ENCODE-Region ENm008. Diese 500 Kb große Region liegt im Anfang von Chromosom 16. In der Abbildung sind drei CNV-*Loci* in einem einzelnen Individuum zu erkennen. Sie manifestieren sich hier in Form von drei Deletionen (ID_7, ID_8, ID_9).

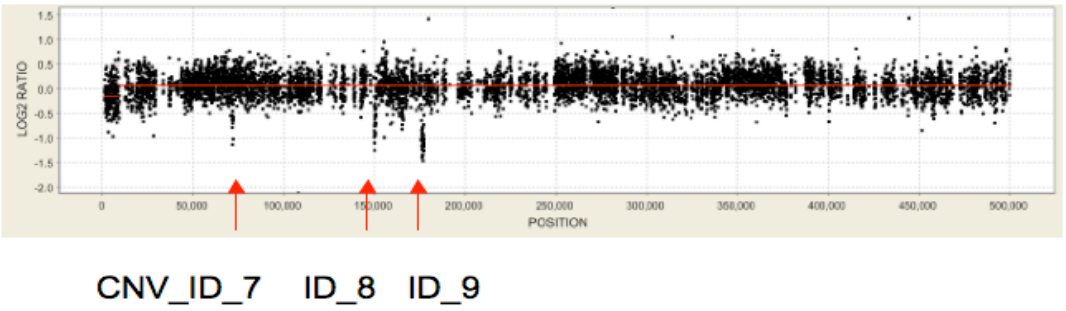
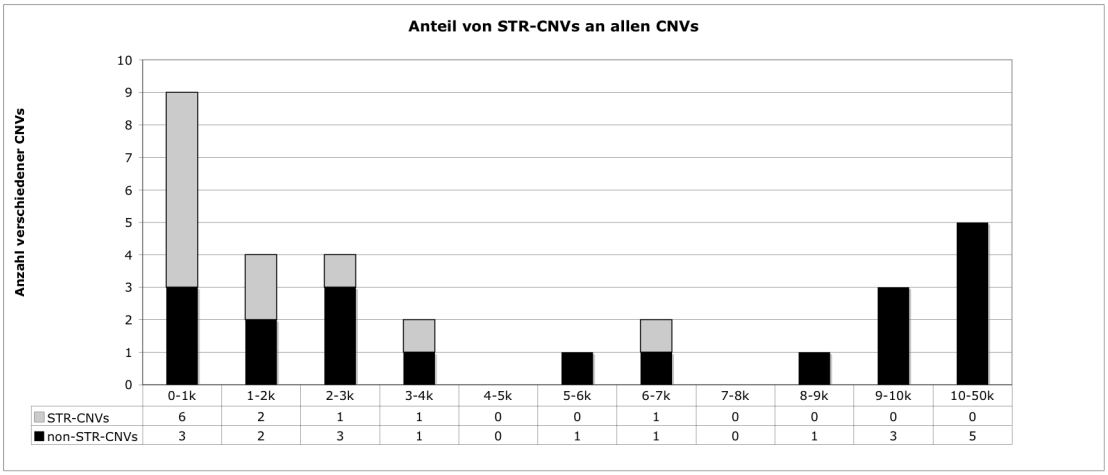



**Abbildung 11: aCGH-Plot von ENCODE-Region ENm008**
Die Encode-Region ENm008 besitzt mehrere CNVs (ID_7, ID_8, ID_9), welche in der Nähe des Telomers liegen.

Die Regionen in der Nähe von Telomeren weisen erfahrungsgemäß eine große Anzahl dieser STRs auf. Diese hypervariablen CNVs haben eine weitere Gemeinsamkeit: eine zugrunde liegende DNA-Sequenz. Die elf CNVs, die auf STRs basieren, sind insgesamt für etwa 20000 variable Basenpaare in den ENCODE-Regionen verantwortlich. Abbildung 12 zeigt die Größenverteilung aller CNVs mit Hinweis darauf, ob eine STR-Sequenz vorliegt. Wie in Abbildung 5 sieht man, dass die Mehrheit der CNVs kleiner als 10 Kb ist, davon sind neun CNVs < 1 Kb. Lediglich fünf CNVs sind größer als 10 Kb. STR-CNVs sind größtenteils < 4 Kb und machen etwa die Hälfte aller CNVs unter 10 Kb Größe aus.




| | 0-1k | 1-2k | 2-3k | 3-4k | 4-5k | 5-6k | 6-7k | 7-8k | 8-9k | 9-10k | 10-50k |
|---|---|---|---|---|---|---|---|---|---|---|---|
| STR-CNVs | 6 | 2 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| non-STR-CNVs | 3 | 2 | 3 | 1 | 0 | 1 | 1 | 0 | 1 | 3 | 5 |

**Abbildung 12: Anteil der STR-CNVs an allen CNVs**
Die Mehrheit der CNVs ist kleiner als 4 Kb, davon neun CNVs < 1 Kb. Die Hälfte der STR-CNVs ist kleiner als 1 Kb.

## 4.2 Validierung der Vorhersagen

Um meine auf *Array*-Daten basierenden Vorhersagen zu validieren, habe ich drei verschiedene Wege beschritten. Ich habe fünf verschiedene *Loci* in jeweils 36 Individuen mittels qPCR und neun weitere *Loci* mit Hilfe der konventionellen PCR untersucht. Außerdem habe ich die CNV-Annotationen in der neueren Literatur zum Vergleich herangezogen.

### 4.2.1 Validierung mittels qPCR

Zur Validierung meiner *Microarray*-Daten habe ich fünf CNV-*Loci* ausgewählt und dann die dort vorliegende Kopienzahl mittels qPCR bestimmt (Tabelle 8). Die ausgesuchten CNVs mussten eine Größe von mindestens 500 bp haben und wurden ausgeschlossen, wenn die zugrunde liegende Sequenz aus STRs bestand.

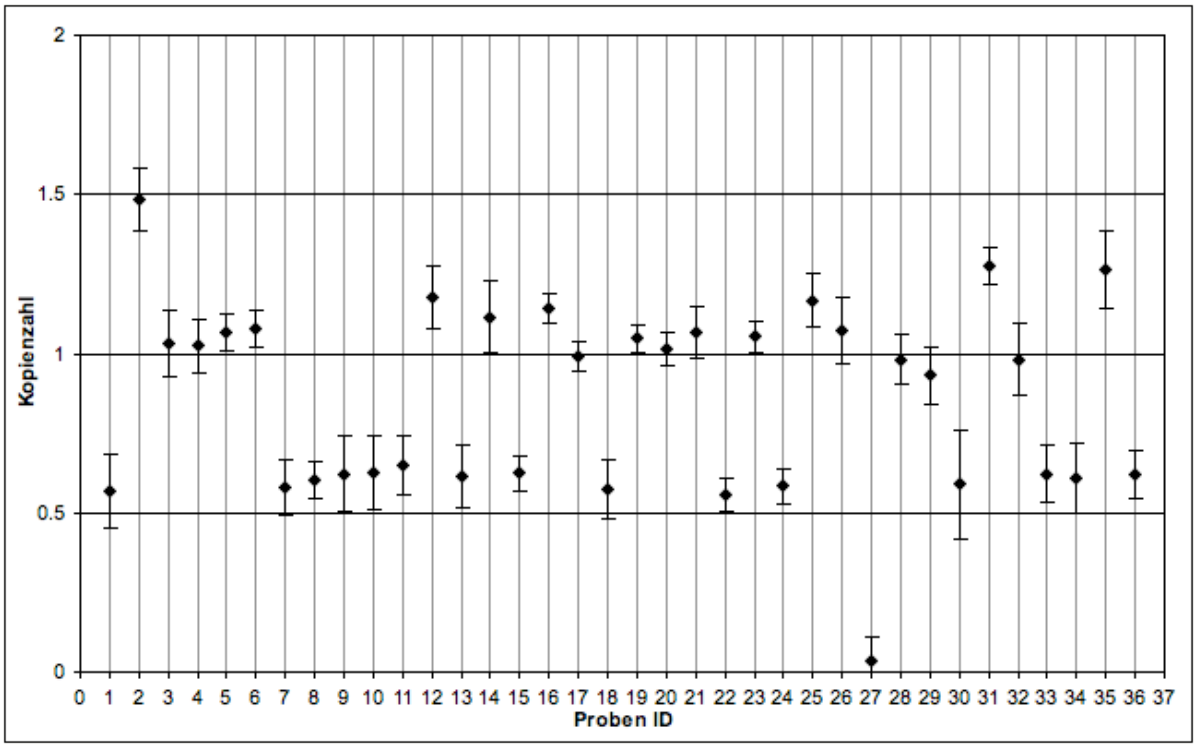
| CNV_ID | Region | Chromosome | Start | Ende | Größe (in bp) |
|---|---|---|---|---|---|
| 10 | ENm009 | chr11 | 4923917 | 4933617 | 9700 |
| 11 | ENm009 | chr11 | 4976629 | 4977237 | 608 |
| 13 | ENm009 | chr11 | 5478760 | 5479520 | 760 |
| 22 | ENm013 | chr7 | 89648606 | 89650354 | 1748 |
| 23 | ENm014 | chr7 | 126563307 | 126568627 | 5320 |

**Tabelle 8: Liste der mit qPCR untersuchten CNV-*Loci***
Eigenschaften der fünf ausgewählten CNVs (ID 10, 11, 13, 22, 23): ENCODE-Region, Chromosom, Start- und Endpunkt sowie Größe des untersuchten CNVs.

Die Abbildungen 13-17 zeigen die Ergebnisse der fünf mittels qPCR untersuchten *Loci*. Alle fünf zeigen Abweichungen von der “normalen” Kopienzahl und bestätigen damit, dass es sich um einen CNV-*Locus* handelt. Die bestimmten Kopienzahlen stimmen mehrheitlich mit den durch die *Arrays* vorhergesagten überein. Die ausgewählten fünf *Loci* liegen innnerhalb von Genen. *Locus* ID_10 betrifft zwei Exone von Olfaktorischen Rezeptorgenen, *Locus* ID_11 ein Exon eines Olfaktorischen Rezeptorgens und die übrigen drei *Loci* liegen in Introns anderer Gene.

Abbildung 13 zeigt die qPCR-Ergebnisse des *Locus* mit der ID_10. Dieser ca. 9700 bp große *Locus* zeigt ein heterogenes Verteilungsmuster, wobei, von zwei Ausnahmen abgesehen, jeweils etwa die Hälfte der Probanden eine “normale” Kopienzahl hat oder eine heterozygote Deletion aufweist. Proband 2 zeigt eine Zunahme der Kopienzahl, Proband 27 eine homozygote Deletion und die Probanden 31 und 35 sind nicht eindeutig in ihrer Kopienzahl. Abbildung 14 für den nur 608 bp großen *Locus* ID_11 zeigt ein homogenes Verteilungsmuster. Bis auf drei Probanden sind alle anderen homozygot für den nicht deletierten Zustand. Die Probanden 5, 9 und 28 zeigen eine heterozygote Deletion. Abbildung 15 zeigt überwiegend homozygote Individuen ohne Deletion. Acht der 36 Probanden zeigen eine heterozygote Deletion für *Locus* ID_13 mit 760 bp. In Abbildung 16 sieht man mehr Variation. 14 Individuen sind heterozygot, neun haben eine homozygote Deletion und 12 zeigen keine Deletion. Individuum 3 hat eine Kopienzahlzunahme. Abbildung 17 zeigt zwei Probanden mit heterozygoter Deletion und die übrigen Probanden im homozygoten nicht-deletierten Zustand. Der *Locus* ID_23 ist ca. 5320 bp groß.

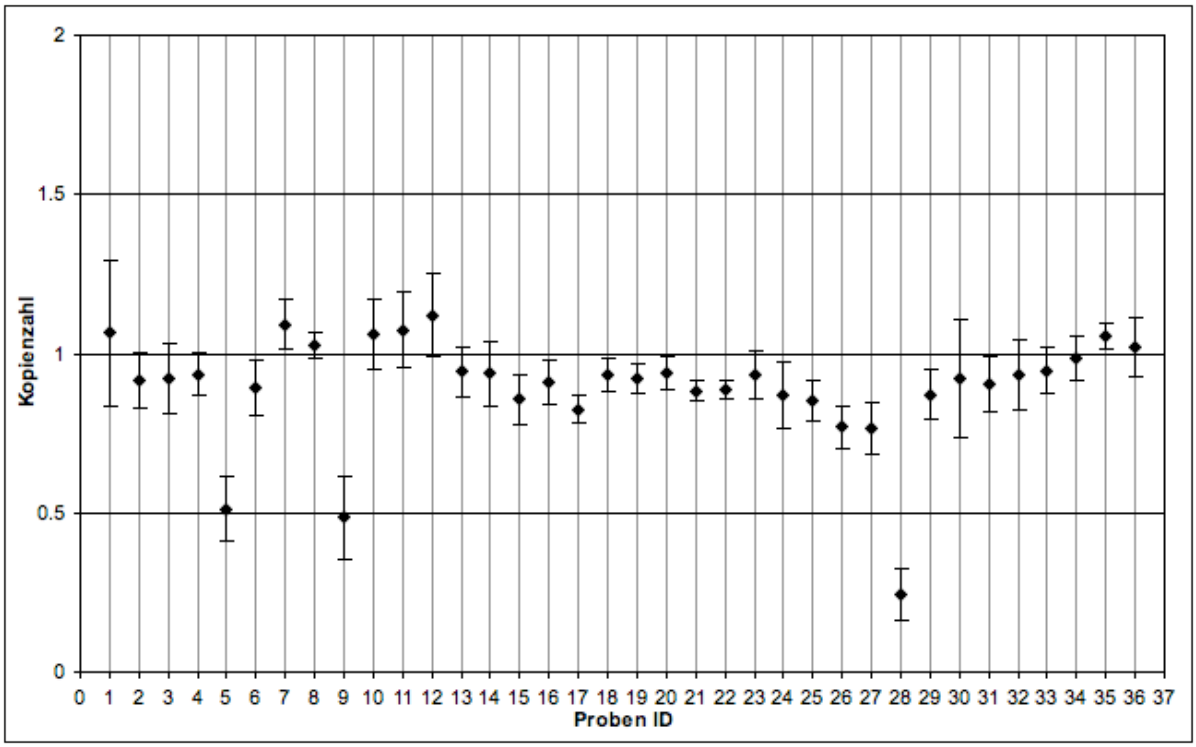
**CNV_ID_10**




| Proben ID | Kopienzahl | STDEV | Proben ID | Kopienzahl | STDEV |
|---|---|---|---|---|---|
| 1 | 0,57 | 0,11 | 19 | 1,05 | 0,04 |
| 2 | 1,48 | 0,10 | 20 | 1,02 | 0,05 |
| 3 | 1,03 | 0,10 | 21 | 1,07 | 0,08 |
| 4 | 1,02 | 0,08 | 22 | 0,56 | 0,05 |
| 5 | 1,07 | 0,06 | 23 | 1,05 | 0,05 |
| 6 | 1,08 | 0,06 | 24 | 0,58 | 0,05 |
| 7 | 0,58 | 0,09 | 25 | 1,17 | 0,08 |
| 8 | 0,60 | 0,06 | 26 | 1,07 | 0,11 |
| 9 | 0,62 | 0,12 | 27 | 0,04 | 0,07 |
| 10 | 0,63 | 0,11 | 28 | 0,98 | 0,08 |
| 11 | 0,65 | 0,09 | 29 | 0,93 | 0,09 |
| 12 | 1,18 | 0,10 | 30 | 0,59 | 0,17 |
| 13 | 0,62 | 0,10 | 31 | 1,28 | 0,06 |
| 14 | 1,12 | 0,11 | 32 | 0,98 | 0,11 |
| 15 | 0,62 | 0,06 | 33 | 0,62 | 0,09 |
| 16 | 1,14 | 0,05 | 34 | 0,61 | 0,11 |
| 17 | 0,99 | 0,05 | 35 | 1,26 | 0,12 |
| 18 | 0,57 | 0,09 | 36 | 0,62 | 0,07 |

**Abbildung 13: qPCR-Ergebnisse für *Locus* ID_10**
Oberer Teil: real-time-PCR Plot für 36 Test-Individuen
Unterer Teil: bestimmte Kopienzahl mit Standardabweichung

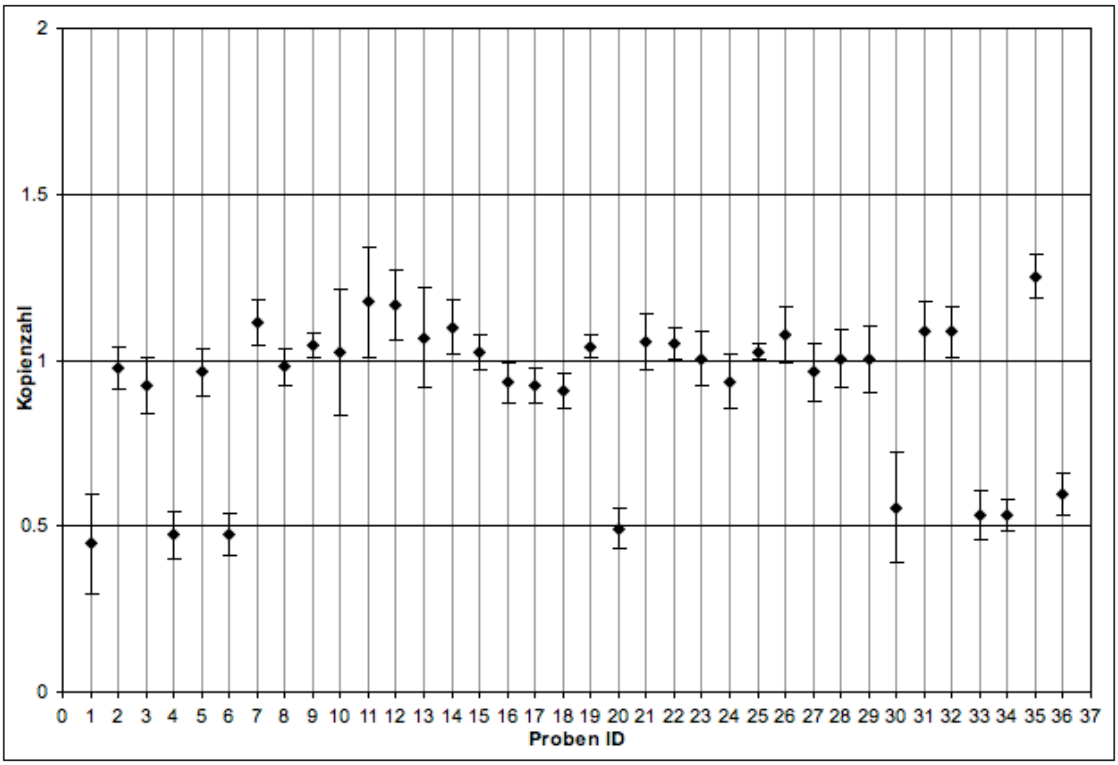
**CNV_ID_11**




| Proben ID | Kopienzahl | STDEV | Proben ID | Kopienzahl | STDEV |
|---|---|---|---|---|---|
| 1 | 1,07 | 0,23 | 19 | 0,92 | 0,05 |
| 2 | 0,92 | 0,09 | 20 | 0,94 | 0,05 |
| 3 | 0,92 | 0,11 | 21 | 0,88 | 0,03 |
| 4 | 0,94 | 0,07 | 22 | 0,89 | 0,03 |
| 5 | 0,51 | 0,10 | 23 | 0,93 | 0,08 |
| 6 | 0,89 | 0,09 | 24 | 0,87 | 0,10 |
| 7 | 1,09 | 0,08 | 25 | 0,85 | 0,06 |
| 8 | 1,03 | 0,04 | 26 | 0,77 | 0,07 |
| 9 | 0,49 | 0,13 | 27 | 0,76 | 0,08 |
| 10 | 1,06 | 0,11 | 28 | 0,24 | 0,08 |
| 11 | 1,08 | 0,12 | 29 | 0,87 | 0,08 |
| 12 | 1,12 | 0,13 | 30 | 0,92 | 0,19 |
| 13 | 0,94 | 0,08 | 31 | 0,90 | 0,09 |
| 14 | 0,94 | 0,10 | 32 | 0,93 | 0,11 |
| 15 | 0,86 | 0,08 | 33 | 0,95 | 0,07 |
| 16 | 0,91 | 0,07 | 34 | 0,98 | 0,07 |
| 17 | 0,83 | 0,04 | 35 | 1,06 | 0,04 |
| 18 | 0,93 | 0,05 | 36 | 1,02 | 0,10 |

**Abbildung 14: qPCR-Ergebnisse für *Locus* ID_11**
Oberer Teil: real-time-PCR Plot für 36 Test-Individuen
Unterer Teil: bestimmte Kopienzahl mit Standardabweichung

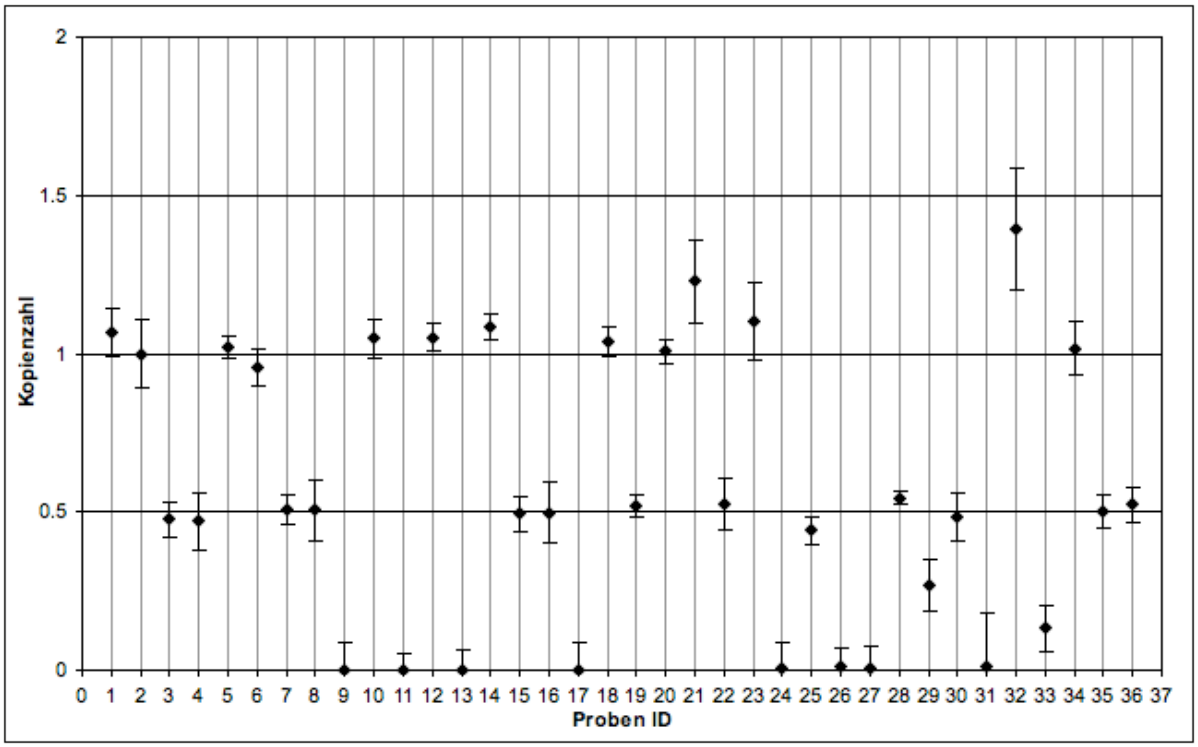
**CNV_ID_13**




| Proben ID | Kopienzahl | STDEV | Proben ID | Kopienzahl | STDEV |
|---|---|---|---|---|---|
| 1 | 0,45 | 0,15 | 19 | 1,04 | 0,04 |
| 2 | 0,98 | 0,06 | 20 | 0,49 | 0,06 |
| 3 | 0,93 | 0,08 | 21 | 1,05 | 0,09 |
| 4 | 0,47 | 0,07 | 22 | 1,05 | 0,05 |
| 5 | 0,97 | 0,07 | 23 | 1,00 | 0,08 |
| 6 | 0,48 | 0,06 | 24 | 0,94 | 0,08 |
| 7 | 1,12 | 0,07 | 25 | 1,03 | 0,02 |
| 8 | 0,98 | 0,05 | 26 | 1,08 | 0,08 |
| 9 | 1,05 | 0,04 | 27 | 0,96 | 0,09 |
| 10 | 1,02 | 0,19 | 28 | 1,00 | 0,09 |
| 11 | 1,18 | 0,17 | 29 | 1,00 | 0,10 |
| 12 | 1,16 | 0,11 | 30 | 0,55 | 0,17 |
| 13 | 1,07 | 0,15 | 31 | 1,09 | 0,09 |
| 14 | 1,10 | 0,08 | 32 | 1,09 | 0,08 |
| 15 | 1,02 | 0,05 | 33 | 0,53 | 0,07 |
| 16 | 0,93 | 0,06 | 34 | 0,53 | 0,05 |
| 17 | 0,92 | 0,05 | 35 | 1,25 | 0,07 |
| 18 | 0,91 | 0,05 | 36 | 0,59 | 0,06 |

**Abbildung 15: qPCR-Ergebnisse für *Locus* ID_13**
Oberer Teil: real-time-PCR Plot für 36 Test-Individuen
Unterer Teil: bestimmte Kopienzahl mit Standardabweichung

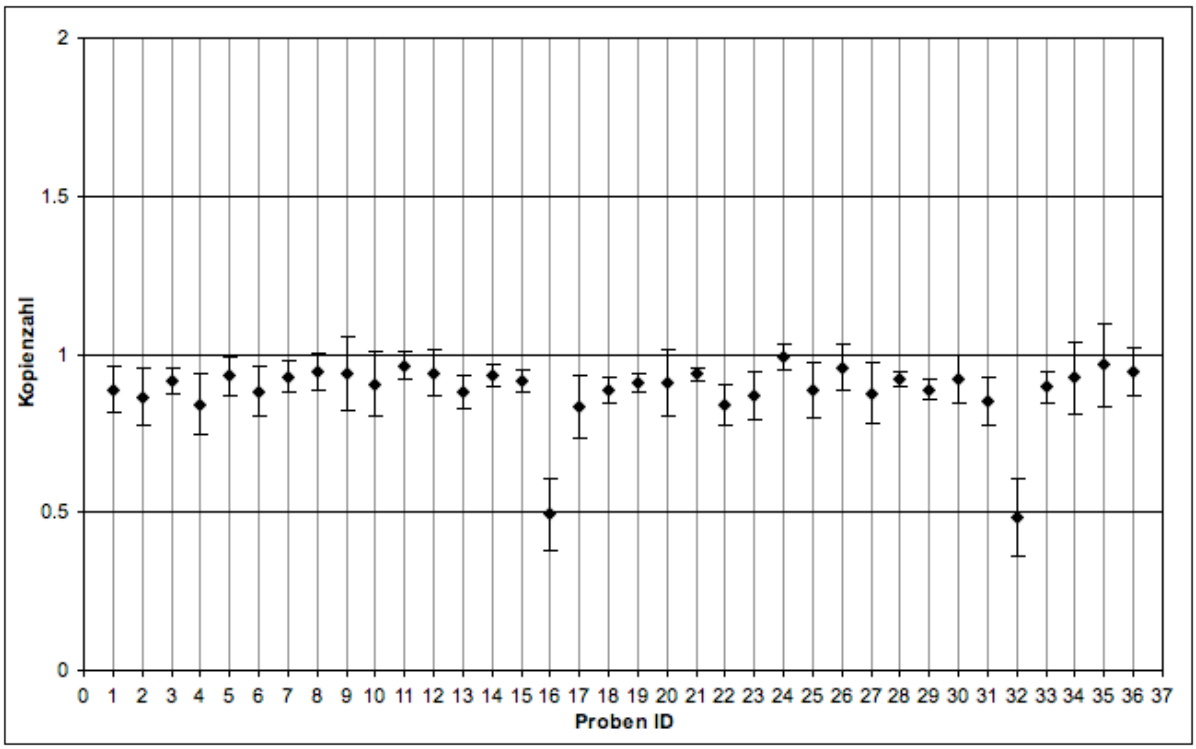
**CNV_ID_22**




| Proben ID | Kopienzahl | STDEV | Proben ID | Kopienzahl | STDEV |
|---|---|---|---|---|---|
| 1 | 1,07 | 0,08 | 19 | 0,52 | 0,03 |
| 2 | 1,00 | 0,11 | 20 | 1,01 | 0,04 |
| 3 | 0,48 | 0,06 | 21 | 1,23 | 0,13 |
| 4 | 0,47 | 0,09 | 22 | 0,52 | 0,08 |
| 5 | 1,02 | 0,04 | 23 | 1,10 | 0,12 |
| 6 | 0,96 | 0,06 | 24 | 0,00 | 0,08 |
| 7 | 0,51 | 0,04 | 25 | 0,44 | 0,04 |
| 8 | 0,51 | 0,10 | 26 | 0,01 | 0,06 |
| 9 | 0,00 | 0,09 | 27 | 0,01 | 0,07 |
| 10 | 1,05 | 0,06 | 28 | 0,54 | 0,02 |
| 11 | 0,00 | 0,05 | 29 | 0,27 | 0,08 |
| 12 | 1,05 | 0,04 | 30 | 0,48 | 0,08 |
| 13 | 0,00 | 0,06 | 31 | 0,01 | 0,17 |
| 14 | 1,09 | 0,04 | 32 | 1,39 | 0,19 |
| 15 | 0,49 | 0,05 | 33 | 0,13 | 0,07 |
| 16 | 0,50 | 0,09 | 34 | 1,02 | 0,09 |
| 17 | 0,00 | 0,09 | 35 | 0,50 | 0,05 |
| 18 | 1,04 | 0,05 | 36 | 0,52 | 0,05 |

**Abbildung 16: qPCR-Ergebnisse für *Locus* ID_22**
Oberer Teil: real-time-PCR Plot für 36 Test-Individuen
Unterer Teil: bestimmte Kopienzahl mit Standardabweichung

**CNV_ID_23**




| Proben ID | Kopienzahl | STDEV | Proben ID | Kopienzahl | STDEV |
|---|---|---|---|---|---|
| 1 | 0,89 | 0,07 | 19 | 0,91 | 0,03 |
| 2 | 0,87 | 0,09 | 20 | 0,91 | 0,10 |
| 3 | 0,92 | 0,04 | 21 | 0,94 | 0,02 |
| 4 | 0,84 | 0,10 | 22 | 0,84 | 0,06 |
| 5 | 0,93 | 0,06 | 23 | 0,87 | 0,07 |
| 6 | 0,88 | 0,08 | 24 | 0,99 | 0,04 |
| 7 | 0,93 | 0,05 | 25 | 0,89 | 0,09 |
| 8 | 0,95 | 0,06 | 26 | 0,96 | 0,07 |
| 9 | 0,94 | 0,12 | 27 | 0,88 | 0,10 |
| 10 | 0,91 | 0,10 | 28 | 0,92 | 0,03 |
| 11 | 0,96 | 0,05 | 29 | 0,89 | 0,03 |
| 12 | 0,94 | 0,07 | 30 | 0,92 | 0,07 |
| 13 | 0,88 | 0,05 | 31 | 0,85 | 0,07 |
| 14 | 0,93 | 0,04 | 32 | 0,48 | 0,12 |
| 15 | 0,91 | 0,04 | 33 | 0,90 | 0,05 |
| 16 | 0,49 | 0,11 | 34 | 0,92 | 0,12 |
| 17 | 0,83 | 0,10 | 35 | 0,97 | 0,13 |
| 18 | 0,89 | 0,04 | 36 | 0,94 | 0,08 |

**Abbildung 17: qPCR-Ergebnisse für *Locus* ID_23**
Oberer Teil: real-time-PCR Plot für 36 Test-Individuen
Unterer Teil: bestimmte Kopienzahl mit Standardabweichung

### 4.2.2 Validierungen mittels konventioneller PCR

Ich habe versucht, neun CNV-*Loci* mit Hilfe der konventionellen PCR zu validieren. Sieben dieser Validierungen waren positiv (s. Tabelle 9). Dazu wurden mit Hilfe von für die jeweiligen CNV-*Loci* spezifischen Primerpaaren die entsprechenden CNV-*Loci* in den verschiedenen Individuen durch konventionelle PCR amplifiziert. Die PCR-Produkte wurden gelelektrophoretisch aufgetrennt und mit den korrespondieren aCGH-Plots verglichen. Abbildung 18 verdeutlicht dies exemplarisch für die ENCODE-Region ENm007, welche sich über eine Region von etwa einer Mb auf Chromosom 19 erstreckt. Es wird schematisch das für *Locus* ID_4 spezifische Primerpaar (Abbildung 18, A) und die gelelektrophoretisch aufgetrennten PCR-Produkte von *Locus* ID_4 der 24 verschiedenen Test-Individuen (Abbildung 18, B) gezeigt. Des Weiteren sieht man die aCGH-Plots von Test-Individuum 7 und 22 für die ENCODE-Region ENm007 (Abbildung 18, C+D). Beide Plots zeigen in *Locus* ID_4 eine Deletion, wobei Test-Individum 22 (Abbildung 18, C) einen stärkeren Abfall der Intensität aufweist, was einer homozygoten Deletion entspricht. Individuum 7 (Abbildung 18, D) zeigt ein weniger abfallendes Signal, was für eine heterozygote Deletion spricht. Dieses Ergebnis wird durch die Ergebnisse der konventionellen PCR bestätigt. So liefert z. B. die konventionelle PCR von *Locus* ID_4 in Individuum 7 zwei verschieden große Banden, während sie in Individuum 22 nur zu einer Bande führt (Abbildung 18, B).

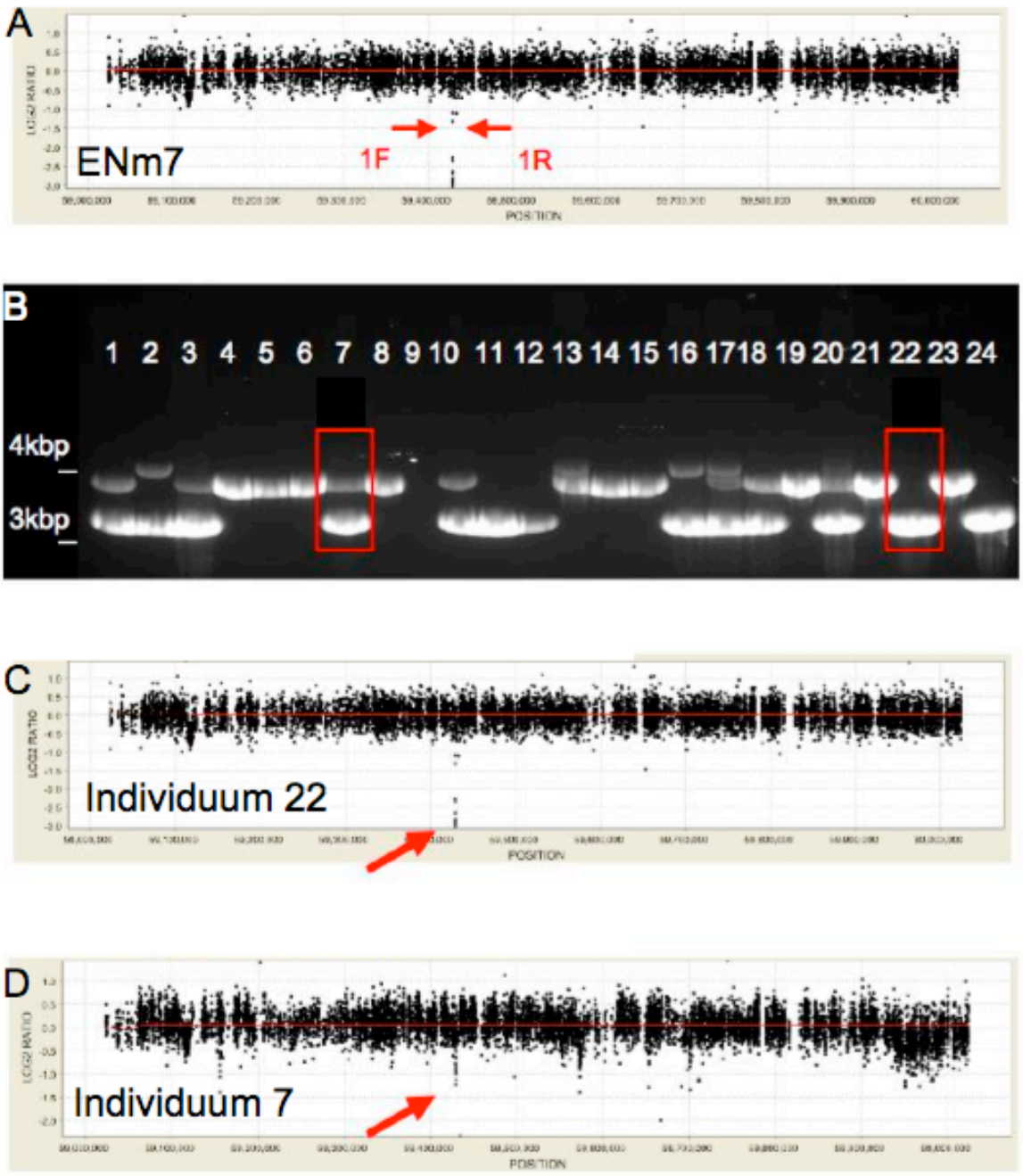



**Abbildung 18: Vergleich der aCGH- mit den PCR-Ergebnissen**
A) Schematische Darstellung der ENCODE-Region ENm007 mit dem für *Locus* ID_4 spezifischen Primerpaar (rote Pfeile). B) Ergebnisse der konventionellen PCR von *Locus* ID_4 von 24 verschiedenen Individuen. Individuum 7 zeigt eine heterozygote, Individuum 22 eine homozygote Deletion (rote Kästchen). C,D) aCGH-Plots der ENCODE-Region ENm007 von Individuum 22 und Individuum 7. Beide zeigen einen Abfall der Intensität in *Locus* ID_4 (C+D, rote Pfeile).

### 4.2.3 Validierung mittels Vergleich mit aus der Literatur bekannten CNVs

Ich habe meine Ergebnisse auch mit den Daten aus der *Database of Genomic Variation* und zusätzlich mit den Daten zweier aktueller wissenschaftlicher Artikel [41, 42] verglichen. Da die meisten bekannten CNV-Annotationen in der DGV sehr groß sind, berücksichtigte ich nur annotierte CNVs, die nicht mehr als doppelt so groß wie meine jeweilige Vorhersage sind. War das annotierte CNV hingegen bis zu 75 % kleiner, habe ich es in den Vergleich mit eingeschlossen, da viele dieser CNVs mit Sequenzvergleichen gefunden wurden, die eine höhere Auflösung als meine Methode haben. Wenn ein CNV diese Kriterien erfüllte, habe ich es als validiert angesehen, da die Wahrscheinlichkeit einer zufällig falsch positiven Übereinstimmung mit einer anderen Methode als sehr gering angesehen werden kann.

Für 10 von 31 meiner CNVs habe ich korrespondierende CNVs in der Literatur gefunden. Sechs davon sind von mir auch durch PCRs validiert worden, d. h. ich habe sechs bestätigte und vier zusätzliche Validierungen durch den Vergleich mit anderen Studien.

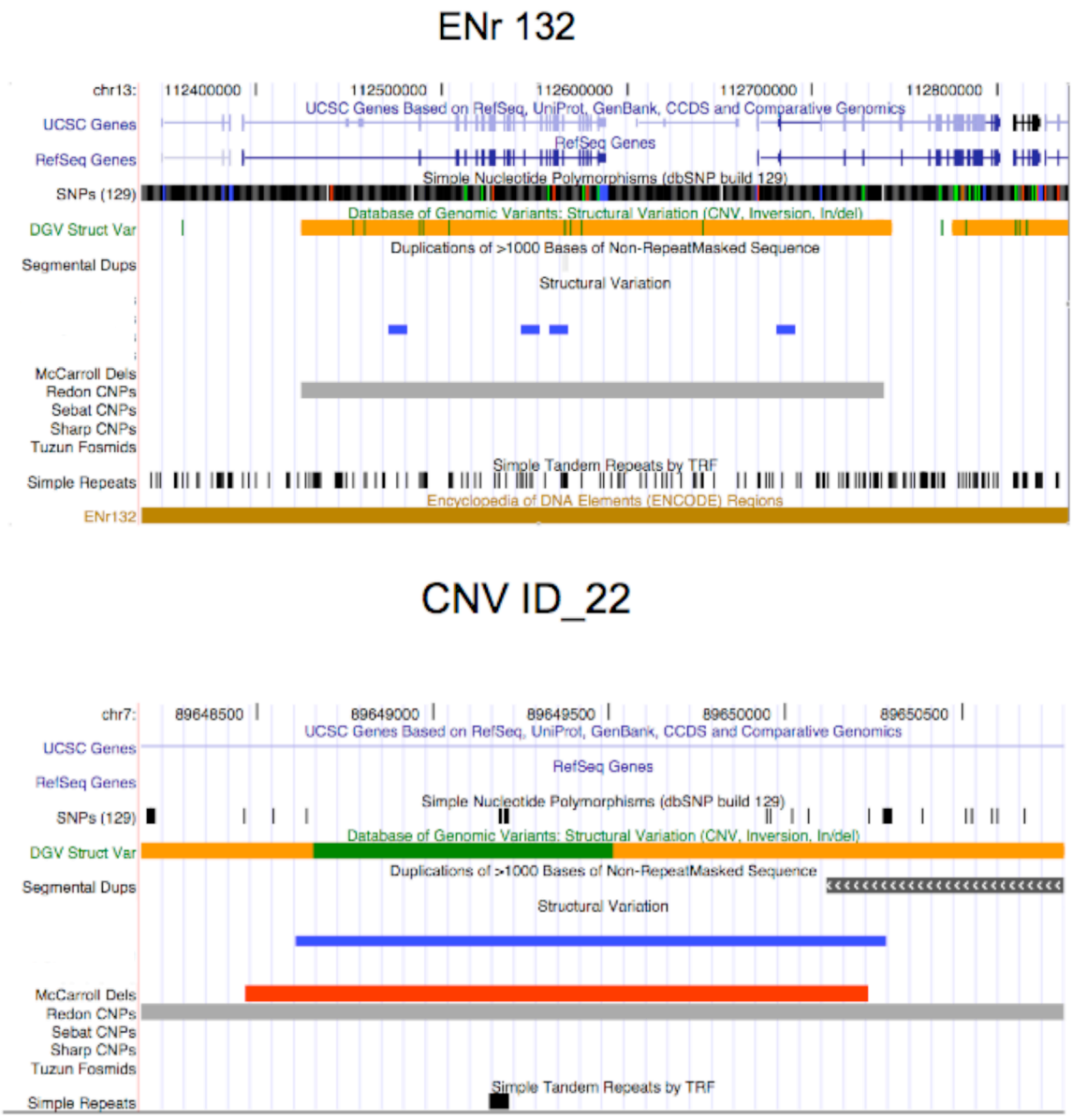



**Abbildung 19: Vergleich von aCGH-Ergebnissen mit in aktueller Literatur beschriebenen CNVs**
Oberer Teil: ENr132. Zu sehen ist ein UCSC-Browser Plot von ENCODE-Region ENr132. Es sind die annotierten Gene für diese Region zu sehen und auch die in der DGV eingetragenen CNVs. Über zwei Drittel der gesamten Region erstreckt sich ein CNV, das von Redon et al. annotiert wurde (grauer Balken). Darin liegen vier meiner Vorhersagen, die sehr viel kleiner sind (blaue Balken).
Unterer Teil: CNV ID_22: Zu sehen ist eine Darstellung von CNV ID_22. Meine Vorhersage (blauer Balken) stimmt nahezu mit den Einträgen in der DGV (grüner und roter Balken) überein. Das CNV gilt als validiert.

### 4.2.4 Zusammenfassung der Validierungsergebnisse

Ich habe versucht, 13 von 31 in meiner Studie gefundenen CNVs mit qPCR oder konventioneller PCR zu validieren. Zwei dieser potentiellen CNVs ließen sich nicht validieren und werden daher als falsch positive Vorhersage gewertet. Damit ergibt sich für die angewandte Methode der aCGH eine Falsch-Positiv-Rate von etwa 15%. Der Vergleich mit in der aktuellen wissenschaftlichen Literatur beschriebenen CNVs bestätigt 10 aus den 31 CNVs. Die im Rahmen der Validierungsexperimente untersuchten CNV-Loci sind in Tabelle 9 zusammengefasst.

| Nr. | Region | qPCR | PCR | Literatur |
|---|---|---|---|---|
| 1 | ENm004 | | | 0 |
| 2 | ENm005 | | | 0 |
| 3 | ENm007 | | 1 | 1 |
| 4 | ENm007 | | 1 | 1 |
| 5 | ENm007 | | | 0 |
| 6 | ENm008 | | 1 | 1 |
| 7 | ENm008 | | 1 | 0 |
| 8 | ENm008 | | 1 | 0 |
| 9 | ENm008 | | | 0 |
| 10 | ENm009 | 1 | | 1 |
| 11 | ENm009 | 1 | | 0 |
| 12 | ENm009 | | | 0 |
| 13 | ENm009 | 1 | | 1 |
| 14 | ENm011 | | | 0 |
| 15 | ENm011 | | | 0 |
| 16 | ENm011 | | | 0 |
| 17 | ENm011 | | | 0 |
| 18 | ENm011 | | | 0 |
| 19 | ENm011 | | | 1 |
| 20 | ENm011 | | | 0 |
| 21 | ENm011 | | | 0 |
| 22 | ENm013 | 1 | 1 | 1 |
| 23 | ENm014 | 1 | | 1 |
| 24 | ENm014 | | | 0 |
| 25 | ENr132 | | 1 | 1 |
| 26 | ENr132 | | | 0 |
| 27 | ENr132 | | | 1 |
| 28 | ENr132 | | | 0 |
| 29 | ENr231 | | | 0 |
| 30 | ENr313 | | 1 | 0 |
| 31 | ENr332 | | 1 | 0 |
| | total | 5 | 9 | 10 |

**Tabelle 9: Zusammenfassung der Validierungsergebnisse**
Sämtliche CNVs sowie die Methoden, mit denen sie validiert wurden. Eine “1” steht für validierte CNVs, leere Kästchen für nicht untersuchte CNVs und eine “0” für keine Übereinstimmnug mit der Literatur. Rote Kästchen zeigen negative Validierungen.

# 5 Diskussion

## 5.1 Anzahl und Größe von CNVs

Ich habe in einem hochauflösenden aCGH-Ansatz 1 Prozent des menschlichen Genoms auf CNVs hin untersucht. Dabei habe ich 31 verschiedene CNV-*Loci* gefunden, die ca. 200 Kb genomischer Sequenz umfassen. Auf das gesamte menschliche Genom hochgerechnet ergibt das 20 Mb oder 0,6 Prozent an Sequenz, die variabel in ihrer Kopienzahl sein kann. Das ist eine deutliche Reduktion der bisher vorherrschenden Schätzungen, die von 12% oder 360 Mb an Variabilität ausgehen [31], um etwa 90% und steht im Einklang mit neueren Arbeiten, die diese Zahl ebenfalls nach unten korrigieren. Diese gehen ebenfalls von einer Reduktion um 80-90% aus und bestätigen damit meine Ergebnisse [41, 42].

Für die in meiner Studie untersuchten Regionen des menschlichen Genoms werden bisher in der aktuellen Literatur nur 10 CNVs beschrieben, die mit denen in dieser Studie entdeckten übereinstimmen. Die restlichen 21 in dieser Arbeit beschriebenen CNVs konnten also mit den in anderen Studien angewandten Methoden nicht detektiert werden.

Auf das gesamte menschliche Genom hochgerechnet erwarte ich ca. 3100 mögliche CNVs, während beide der oben genannten Studien von ca. 1400 CNVs genomweit ausgehen.

Die große Mehrheit der von mir gefundenen CNVs ist kleiner als 10 Kb. Dieses bestätigt die Ergebnisse verschiedener Studien, die entweder neuere *Array*-Methoden oder verschiedene Sequenziertechniken benutzt haben [32] [43]. Das Maximum dieser Verteilung liegt bei einer Größe kleiner 1 Kb, während andere Studien ihre Maxima zwischen 3-9 Kb haben.

Dies macht deutlich, dass die meisten der älteren Annotationen/Studien die Größe von CNVs überschätzt haben. Der wahrscheinlichste Grund dafür ist, dass die älteren Experimente oft mit BAC-*Arrays* durchgeführt wurden, deren Auflösung durch die Größe der einzelnen BAC-Klone (ca. 150 Kb) limitiert wird. Somit können die genauen Grenzen eines CNVs nicht bestimmt werden, und mehrere kleinere CNVs werden teilweise als ein größeres beschrieben [44]. Das ist eine gute Erklärung dafür, dass ich in meiner Studie mehr verschiedene CNVs gefunden

habe, bei gleichzeitiger Abnahme der Gesamtzahl betroffener Basenpaare.
Diese beiden Punkte bestätigen die dieser Arbeit zugrunde liegende Hypothese, dass bisherige Methoden zur CNV-Detektion einerseits eine große Zahl von Kopienzahlveränderungen im menschlichen Genom nicht detektieren konnten und andererseits aufgrund mangelnder Auflösung dieser Methoden die Menge an kopienzahlvariabler Basenpaare stark überschätzt wurde.

## 5.2 Vergleich mit anderen Studien

Ein Ziel meiner Studie war es, die Möglichkeiten von HR-CGH mit einem überlappenden Oligomerpfad im Vergleich zu anderen Methoden zu evaluieren und die von mir gefundenen CNVs mit den in der Literatur beschriebenen zu vergleichen. Dieser Vergleich kann problematisch sein, da sehr viele Variablen zu berücksichtigen sind, z. B. verschiedene Probanden-DNA, unterschiedliche Plattformen, Algorithmen zur Datenanalyse, Ausschlusskriterien etc.
Eine Fülle von CNVs ist in der *Database of Genomic Variation* beschrieben. Da diese Daten teilweise die Ausdehnung der betreffenden Region überschätzen, verwundert es nicht, dass von meinen 31 CNVs 24 innerhalb schon beschriebener CNVs lagen. Deshalb habe ich die Kriterien für die Übereinstimmung von zwei CNVs so angepasst, dass die bekannten CNVs nicht mehr als doppelt so groß bzw. 75 % kleiner als meine Vorhersagen sein durften (s. Ergebnisse). Danach stimmten noch zehn CNVs mit zuvor beschriebenen überein.
Besonderes Augenmerk habe ich den Daten von Perry et al. und McCarroll et al. zukommen lassen, da beide Studien neueste Techniken verwenden und eine teilweise Übereinstimmung mit meinem Probensatz besteht. Perry et al. haben in einem aCGH-Assay, das sich auf ca. 2200 vorher bekannte CNV-*Loci* konzentriert hat, 30 HapMap-Individuen untersucht. Drei davon sind auch in meiner Arbeit untersucht worden. Perry et al. finden für diese drei Individuen insgesamt 15 CNV-*Loci* in den ENCODE-Regionen, von denen fünf mit meinen CNV-*Loci* übereinstimmen.
McCarroll et al. haben alle 270 HapMap-Individuen mit einem neuartigen *Array* untersucht, der aCGH und SNP-Genotypisierung miteinander verbindet. Insgesamt haben sie 12 CNVs in den ENCODE-Regionen gefunden. In den Individuen, die

auch in meinem Probensatz untersucht wurden, haben sie vier CNV-*Loci* beschrieben. Diese vier *Loci* finden sich auch in meinen Proben.
Der Vergleich mit den Studien von Perry et al. und McCarroll et al. zeigt teilweise Übereinstimmungen. Für Perry et al. stimmen etwa die Hälfte der CNVs mit meinen überein. Dies ist ein guter Wert, wenn man bedenkt, dass ihre Methode sich auf schon beschriebene CNVs beschränkt. McCarroll et al. beschreiben vier CNVs, die ich auch gefunden habe. Hier ist jedoch die Übereinstimmung geringer; wahrscheinlich deshalb, weil die Methode, die von McCarroll et al. angewandt wurde, ebenfalls eine Verzerrung hin zu schon bekannten kopienzahlvariablen Regionen des Genoms aufweist.

## 5.3 Viele CNVs sind kleiner als 1 Kb

Die bisherigen Grenzen der Auflösung der meisten aCGH- oder SNP-*Arrays* liegt bisher bei etwa 50 Kb; in einigen Fällen für bestimmte Regionen im Genom kann idealerweise eine Auflösung von etwa 2 Kb erreicht werden. Mit HR-CGH können die Bruchpunkte vieler CNVs bis auf wenige hundert Basenpaare genau bestimmt werden, sogar CNVs im Bereich kleiner als 1 Kb können detektiert werden. Dadurch konnte ich mehrere sehr kleine CNVs in meinen Proben nachweisen, die mit anderen Methoden zuvor nicht gefunden werden konnten.
Ein Teil dieser kleinen CNVs hat eine *Simple Tandem Repeat Sequence* (STR) zur Grundlage. Relativ wenige dieser STRs sind als CNVs annotiert. Ich habe dennoch Übereinstimmungen mit Eintragungen in der DGV gefunden, diese stammen jedoch meist aus Studien, die auf Sequenziertechniken basieren [43] (s. Abbildung 20). Ein weiterer möglicher Grund dafür ist, dass sie sich meist in der Nähe des Telomers befinden. Diese Regionen sind in vielen Untersuchungen unterrepräsentiert.

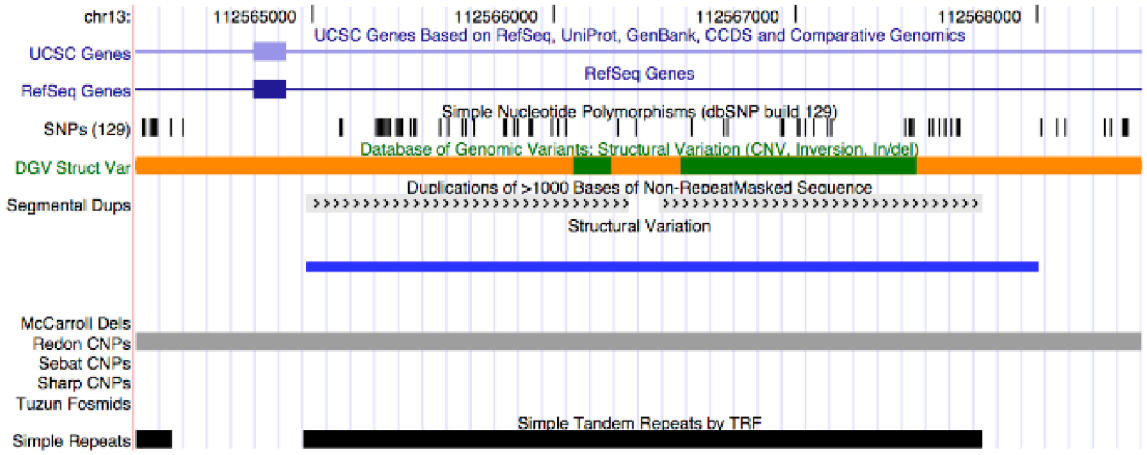



**Abbildung 20: CNV ID_27: STR-CNV und annotiertes InDel**
In Blau zu sehen ist meine Vorhersage für ID_27. Diese überlappt mit einem STR (schwarzer Balken), zwei Segment-Duplikationen und zwei kleineren InDels (grün), die mit einer Sequenziertechnik annotiert wurden.

## 5.4 Verteilung von CNVs

Die verschiedenen ethnischen Gruppen, die ich untersucht habe, zeigen verschiedene Häufigkeiten für die Verteilung von CNVs. Die Gruppe der Yoruber hat den höchsten Durchschnitt an CNVs in den ENCODE-Regionen mit 14 pro Individuum. Dieses Phänomen stimmt mit den Beobachtungen anderer Studien überein und lässt sich mit der „Bottleneck“-Theorie [45] erklären: Die menschliche Population, die Afrika vor ca. 50000 Jahren verlassen hat und von der alle Populationen außerhalb Afrikas abstammen, stellte nur eine relativ kleine Auswahl an Individuen dar und wies daher relativ wenig genetische Diversität auf. Die genetische Diversität zwischen Individuen innerhalb Afrikas hingegen ist deshalb weitaus größer.

Die Beobachtung, dass die CNVs, die in Exons von Genen liegen, überwiegend Gene für Olfaktorische Rezeptoren und die Immunmodulation betreffen, überrascht nicht, da diese Verteilung in der Literatur beschrieben wird und beide Gruppen von Genen von einer flexiblen Kopienzahl profitieren können. Allerdings darf man nicht vergessen, dass diese Zielregionen manuell für das ENCODE-Projekt ausgewählt wurden, da sie von besonderem Interesse waren.

## 5.5 Die Frage der richtigen Kontrolle

Bei aCGH werden eine zu testende DNA und eine Kontroll-DNA miteinander verglichen. Es gibt verschiedene Möglichkeiten, diese Kontrolle zu wählen. Einige Studien haben ein einzelnes Individuum für alle Experimente als Kontrolle. Das Ergebnis bezieht sich dabei auf ein einzelnes Individuum als Referenz. Der Vorteil ist, dass man im Falle einer unterschiedlichen Kopienzahl an einem *Locus* zwischen Test- und Kontroll-DNA einen eindeutigen Wert erhält, d. h ein Vielfaches von 0.5. Der Nachteil ist, dass man, wenn z. B. sowohl Test- als auch Kontroll-DNA dieselbe Deletion aufweisen, kein CNV-Signal sehen würde. Man müsste also alle CNVs eines Kontroll-Individuums vorab kennen. Doch das ist zur Zeit nicht möglich.
Ich habe mich daher dafür entschieden, einen Pool aus 7 Individuen als Kontrolle zu benutzen, um die Möglichkeit zu minimieren, dass Probe und Kontrolle die gleiche Abweichung aufweisen. Der Nachteil ist, dass viele CNVs häufiger sind als zuvor gedacht und diese CNVs auch in meinem Kontroll-Pool vorhanden sind.
Tabelle 5 zeigt alle Ergebnisse für die 31 CNV-*Loci* nach Ethnien geordnet. Eine Auffälligkeit ist, dass die Gruppe der Kaukasier relativ wenig CNVs im Vergleich zu den anderen Gruppen aufweist. Die wahrscheinlichste Erklärung hierfür ist, dass der Kontroll-Pool v. a. von Kaukasiern abstammt und somit die Anzahl an CNVs in den kaukasischen Test-Individuen unterschätzt wird.
Beide Arten der Kontrolle haben also Vor- und Nachteile. Wahrscheinlich wäre in der Zukunft die eleganteste Lösung, eine Referenz-DNA zu verwenden, nachdem diese gegen mehrere hundert Individuen getestet wurde und somit der genaue Genotyp bekannt wäre.

## 5.6 Ausblick für die Methode

Ein Vorzug dieser Methode ist sicher, dass eine sehr hohe Auflösung erreichet werden kann, die teilweise in den Grenzbereich von Sequenziertechniken übergeht (s. Abbildung 20). Auf der anderen Seite ist der Teil des Genoms, den ich hier betrachte, sehr klein. Seit ich meine Studie begonnen habe, hat sich die Technik jedoch fortentwickelt: Inzwischen sind *Arrays* erhältlich, die die fünffache Abdeckung zulassen, und in Zukunft werden *Arrays* erhältlich sein, die das gesamte menschliche Genom mit einem überlappenden Pfad abdecken werden.
Dadurch, dass sich meine Methode nicht auf vorher bekannte CNV-*Loci* beschränkt oder bestimmte Regionen des Genoms wegen fehlender Abdeckung mit SNPs

auslässt, ist es möglich, seltenere CNVs auch *ab initio* aufzufinden. Das ist ein entscheidender Punkt in der humangenetischen Forschung [46].
Welche Bedeutung dieser Technik in der Zukunft zukommen wird, ist schwer zu sagen, besonders unter Berücksichtigung der immer besser und billiger werdenden Sequenziermethoden. Aber vor dem Hintergrund, dass das „1000-Dollar-Genom“ noch nicht in allernächster Zukunft liegt, wird aCGH wohl noch eine Weile einen Platz in der biomedizinischen Forschung einnehmen.

# 6 Zusammenfassung

Seit der ursprünglichen Entdeckung im Jahre 2004 ist immer klarer geworden, dass Kopienzahlveränderungen (CNVs) einen starken Einfluss auf die genetische Prädisposition zu Krankheiten, aber auch auf Variation im gesunden Phänotypen haben. Auch dank verbesserter Methoden werden zunehmend CNVs als Faktoren der molekularen Ätiologie von vielerlei Krankheitsbildern identifiziert. Eine Kartierung aller CNVs im menschlichen Genom ist daher entscheidend für die weitere Erforschung komplexer Erkrankungen und eröffnet neue Ansätze in der Medikamentenentwicklung.

Ich habe mit einer hochauflösenden Methode, Array Comparative Genomic Hybridization (aCGH), die 44 ENCODE-Regionen in 36 Individuen auf CNVs hin untersucht. Dafür habe ich *Microarray Chips* mit 385000 Oligomeren benutzt, was es möglich machte, ein Prozent des menschlichen Genoms mit einem überlappenden Pfad abzudecken. Insgesamt konnte ich so 31 verschiedene CNV-*Loci* mit einer Gesamtausdehnung von ca. 200 Kb finden, von denen der größte Teil mittels quantitativer real-time PCR (qPCR) sowie konventioneller PCR und Datenbankvergleichen validiert werden konnte. Auf das gesamte Genom hochgerechnet entspricht das 20 Mb an Kopienzahl-variabler Sequenz. Damit konnte ich im Einklang mit jüngsten Forschungsergebnissen anderer zeigen, dass die Gesamtzahl an von CNVs betroffenen Basen geringer ist als bislang vermutet, da CNVs zwar zahlreich, aber oft von geringerer Ausdehnung als zuvor angenommen sind. So stellte ich fest, dass die meisten CNV-*Loci* kleiner als 10 Kb sind, mit einem Maximum kleiner als 1 Kb. Im Bereich unter 1 Kb war es mir möglich, eine große Anzahl von CNVs, die sehr variable Sequenzen aufweisen, aufzufinden. Das war bislang lediglich nur mit aufwändigeren Sequenziermethoden möglich. Darüber hinaus konnte ich bestätigen, dass populationspezifische Verteilungsmuster für CNV-*Loci* existieren und sich CNVs überproportional häufig in Olfaktorischen Rezeptorgenen bzw. in Genen, die der Immunmodulation dienen, finden lassen.

# 7 Literaturverzeichnis


1. Antonarakis, S.E., et al., *Chromosome 21 and down syndrome: from genomics to pathophysiology.* Nat Rev Genet, 2004. **5**(10): p. 725-38.
2. Spritz, R.A., et al., *Base substitution in an intervening sequence of a beta+-thalassemic human globin gene.* Proc Natl Acad Sci U S A, 1981. **78**(4): p. 2455-9.
3. *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.
4. *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.
5. Sachidanandam, R., et al., *A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.* Nature, 2001. **409**(6822): p. 928-33.
6. *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-45.
7. Kan, Y.W. and A.M. Dozy, *Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation.* Proc Natl Acad Sci U S A, 1978. **75**(11): p. 5631-5.
8. Lupski, J.R., et al., *DNA duplication associated with Charcot-Marie-Tooth disease type 1A.* Cell, 1991. **66**(2): p. 219-32.
9. Barber, J.C., et al., *Duplication of 8p23.1: a cytogenetic anomaly with no established clinical significance.* J Med Genet, 1998. **35**(6): p. 491-6.
10. Engelen, J.J., et al., *Duplication of chromosome region 8p23.1-->p23.3: a benign variant?* Am J Med Genet, 2000. **91**(1): p. 18-21.
11. Perry, G.H., et al., *Diet and the evolution of human amylase gene copy number variation.* Nat Genet, 2007. **39**(10): p. 1256-60.
12. Hasin, Y., et al., *High-resolution copy-number variation map reflects human olfactory receptor diversity and evolution.* PLoS Genet, 2008. **4**(11): p. e1000249.
13. Young, J.M., et al., *Extensive copy-number variation of the human olfactory receptor gene family.* Am J Hum Genet, 2008. **83**(2): p. 228-42.

14. Gonzalez, E., et al., *The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility.* Science, 2005. **307**(5714): p. 1434-40.
15. Aitman, T.J., et al., *Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans.* Nature, 2006. **439**(7078): p. 851-5.
16. Fanciulli, M., et al., *FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity.* Nat Genet, 2007. **39**(6): p. 721-3.
17. Fellermann, K., et al., *A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon.* Am J Hum Genet, 2006. **79**(3): p. 439-48.
18. Sebat, J., et al., *Strong association of de novo copy number mutations with autism.* Science, 2007. **316**(5823): p. 445-9.
19. Xu, B., et al., *Strong association of de novo copy number mutations with sporadic schizophrenia.* Nat Genet, 2008. **40**(7): p. 880-5.
20. Hollox, E.J., et al., *Psoriasis is associated with increased beta-defensin genomic copy number.* Nat Genet, 2008. **40**(1): p. 23-5.
21. Lupski, J.R., *Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits.* Trends Genet, 1998. **14**(10): p. 417-22.
22. Singleton, A.B., et al., *alpha-Synuclein locus triplication causes Parkinson's disease.* Science, 2003. **302**(5646): p. 841.
23. Rovelet-Lecrux, A., et al., *APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy.* Nat Genet, 2006. **38**(1): p. 24-6.
24. Stranger, B.E., et al., *Relative impact of nucleotide and copy number variation on gene expression phenotypes.* Science, 2007. **315**(5813): p. 848-53.
25. Sebat, J., et al., *Large-scale copy number polymorphism in the human genome.* Science, 2004. **305**(5683): p. 525-8.
26. Iafrate, A.J., et al., *Detection of large-scale variation in the human genome.* Nat Genet, 2004. **36**(9): p. 949-51.
27. Tuzun, E., et al., *Fine-scale structural variation of the human genome.* Nat Genet, 2005. **37**(7): p. 727-32.

28. Conrad, D.F., et al., *A high-resolution survey of deletion polymorphism in the human genome.* Nat Genet, 2006. **38**(1): p. 75-81.
29. Hinds, D.A., et al., *Common deletions and SNPs are in linkage disequilibrium in the human genome.* Nat Genet, 2006. **38**(1): p. 82-5.
30. McCarroll, S.A., et al., *Common deletion polymorphisms in the human genome.* Nat Genet, 2006. **38**(1): p. 86-92.
31. Redon, R., et al., *Global variation in copy number in the human genome.* Nature, 2006. **444**(7118): p. 444-54.
32. Korbel, J.O., et al., *Paired-end mapping reveals extensive structural variation in the human genome.* Science, 2007. **318**(5849): p. 420-6.
33. *The ENCODE (ENCyclopedia Of DNA Elements) Project.* Science, 2004. **306**(5696): p. 636-40.
34. Selzer, R.R., et al., *Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH.* Genes Chromosomes Cancer, 2005. **44**(3): p. 305-19.
35. Urban, A.E., et al., *High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays.* Proc Natl Acad Sci U S A, 2006. **103**(12): p. 4534-9.
36. *http://ccr.coriell.org.*
37. Nuwaysir, E.F., et al., *Gene expression analysis using oligonucleotide arrays produced by maskless photolithography.* Genome Res, 2002. **12**(11): p. 1749-55.
38. Fodor, S.P., et al., *Light-directed, spatially addressable parallel chemical synthesis.* Science, 1991. **251**(4995): p. 767-73.
39. Smit, A.F.A., Hubley, R. & Green, P., *RepeatMasker Open-3.0. 1996-2010 <http://www.repeatmasker.org>.*
40. Pfaffl, M.W., *A new mathematical model for relative quantification in real-time RT-PCR.* Nucleic Acids Res, 2001. **29**(9): p. e45.
41. Perry, G.H., et al., *The fine-scale and complex architecture of human copy-number variation.* Am J Hum Genet, 2008. **82**(3): p. 685-95.
42. McCarroll, S.A., et al., *Integrated detection and population-genetic analysis of SNPs and copy number variation.* Nat Genet, 2008. **40**(10): p. 1166-74.
43. Levy, S., et al., *The diploid genome sequence of an individual human.* PLoS Biol, 2007. **5**(10): p. e254.

44. Carter, N.P., *Methods and strategies for analyzing copy number variation using DNA microarrays.* Nat Genet, 2007. **39**(7 Suppl): p. S16-21.
45. Cavalli-Sforza L. L., P.A., Menozzi P., *History and geography of human genes*. Princeton, NJ: Princeton University Press, 1994.
46. Cohen, J.C., et al., *Multiple rare alleles contribute to low plasma levels of HDL cholesterol.* Science, 2004. **305**(5685): p. 869-72.

# *8 Anhang*

## 8.1 Lebenslauf

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

## 8.2 Selbstständigkeitserklärung

Hiermit erkläre ich, Fabian Grubert, dass die Dissertation mit dem Titel „Strukturvarianten im menschlichen Genom“ von mir selbst und ohne die Hilfe Dritter verfasst wurde, auch in Teilen keine Kopie anderer Arbeiten darstellt und die benutzten Hilfsmittel sowie die Literatur vollständig angegeben sind.

Datum:

Fabian Grubert

## 8.3 Publikationsliste

Stand: 01. Mai 2010


1. Korbel JO*, Urban AE*, **Grubert** F, Du J, Royce TE, Starr P, Zhong G, Emanuel BS, Weissman SM, Snyder M, Gerstein MB, *Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome.* Proc Natl Acad Sci U S A, 2007. **104**(24): p. 10110-5.
2. Korbel JO*, Urban AE*, Affourtit JP*, Godwin B, **Grubert** F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M., *Paired-end mapping reveals extensive structural variation in the human genome.* Science, 2007. **318**(5849): p. 420-6.
3. Pan X, Urban AE, Palejev D, Schulz V, **Grubert** F, Hu Y, Snyder M, Weissman SM., *A procedure for highly specific, sensitive, and unbiased whole-genome amplification.* Proc Natl Acad Sci U S A, 2008. **105**(40): p. 15499-504.
4. Kim PM, Lam HY, Urban AE, Korbel JO, Affourtit JP, **Grubert** F, Chen X, Weissman SM, Snyder M, Gerstein MB., *Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history.* Genome Res, 2008. **18**(12): p. 1865-74.
5. Korbel JO*, Tirosh-Wagner T*, Urban AE*, Chen X, Kasowski M, Dai L, **Grubert** F, Erdman C, Gao MC, Lange K, Sobel EM, Barlow GM, Aylsworth AS, Carpenter NJ, Clark RD, Cohen MY, Doran E, Falik-Zaccai T, Lewin SO, Lott IT, McGillivray BC, Moeschler JB, Pettenati MJ, Pueschel SM, Rao KW, Shaffer LG, Shohat M, Van Riper AJ, Warburton D, Weissman S, Gerstein MB, Snyder M, Korenberg JR., *The genetic architecture of Down syndrome phenotypes revealed by high-resolution analysis of human segmental trisomies.* Proc Natl Acad Sci U S A, 2009. **106**(29): p. 12031-6.
6. Kasowski M*, **Grubert** F*, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE, Hong M, Karczewski KJ,

Huber W, Weissman SM, Gerstein MB, Korbel JO, Snyder M., *Variation in transcription factor binding among humans.* Science, 2010. **328**(5975): p. 232-5.

* Diese Autoren haben zu gleichen Teilen zur Studie beigetragen.

## 8.4 Danksagung

Für die vielfältige Hilfe bei der Erstellung meiner Doktorarbeit möchte ich folgenden Menschen besonders danken:

- Herrn Professor Stefan Mundlos für die Betreuung meiner Arbeit im Rahmen meiner externen Promotion

- Herrn Professor Sherman Weissman für die Möglichkeit, in einer fantastischen Umgebung Einblick in die Grundlagen, und teilweise Feinheiten, der molekularen Genetik zu gewinnen

- Herrn Professor Mike Snyder für die Möglichkeit meine Untersuchungen in seinem Labor fortzusetzen

- Alexander Eckehart Urban für die konzeptionellen Diskussionen bei der Projektplanung und für die Korrektur dieses Manuskriptes

- Jan Korbel für die grundlegende Einführung in die Anwendung verschiedener Software-Pakete

- Yasuo Koga für seine Geduld mit einem absoluten Beginner

- Subhradip Kharmakar und Jin Lian für ihre niemals endende Hilfsbereitschaft

- Der gesamten Weissman AG für ein Arbeitsumfeld, in dem ich mich vom ersten Tag an zu Hause gefühlt habe

- Alexander Kühn für die Betreuung meiner ersten Gehversuche in molekularer Genetik und für die Korrektur dieses Manuskriptes

- Dem Team vom Promotionsbüro der Charité für ermutigenden Zuspruch