WS2017/2018 MSc Modul 7A

„Genomforschung und Sequenzanalyse

- Einführung in Methoden der Bioinformatik- “

AG Hankeln

Trio-Analyse (Exome-Seq)

Mutationen mit Krankheitsrelevanz im Humangenom

More than 3800 genes are (known to be) involved in human disease...

Online Mendelian Inheritance of Man = catalogue of human genes and diseases

OMIM Gene Map Statistics

OMIM Morbid Map Scorecard (Updated November 16th, 2017) :

Total number of phenotypes* for which the molecular basis is known	6,112
Total number of genes with phenotype-causing mutation	3,844

* Phenotypes include (1) single-gene mendelian disorders and traits; (2) susceptibilities to cancer and complex disease (e.g., BRCA1 and familial breast-ovarian cancer susceptibility, 113705.0001, and CFH and macular degeneration, 134370.0008); (3) variations that lead to abnormal but benign laboratory test values ("nondiseases") and blood groups (e.g., lactate dehydrogenase B deficiency, 150100.0001 and ABO blood group system, 110300.0001); and (4) select somatic cell genetic disease (e.g., GNAS and McCune-Albright syndrome, 139320.0008 and IDH1 and glioblastoma multiforme, 147700.0001.)

Distribution of Phenotypes across Genes (Updated November 16th, 2017) :

Number of genes with 1 phenotype	2,631
Number of genes with 2 phenotypes	721
Number of genes with 3 phenotypes	259
Number of genes with 4+ phenotypes	233

Dissected OMIM Morbid Map Scorecard (Updated November 16th, 2017) :

Class of phenotype	Phenotype	Gene *
Single gene disorders and traits	5,072	3,464
Susceptibility to complex disease or infection	698	499
"Nondiseases"	145	115
Somatic cell genetic disease	212	121

*Some genes may be counted more than once because mutations in a gene may cause more than one phenotype and the phenotypes may be of different classes (e.g., activating somatic BRAF mutation underlying cancer, 164757.0001. and germline BRAF mutation in Noonan syndrome, 164757.0022.)

+ 141900

HEMOGLOBIN--BETA LOCUS; HBB

Other entities represented in this entry:

METHEMOGLOBINEMIA, BETA-GLOBIN TYPE, INCLUDED
ERYTHREMIA, BETA-GLOBIN TYPE, INCLUDED

HGNC Approved Gene Symbol: HBB

Cytogenetic location: 11p15.4 ***Genomic coordinates (GRCh38): 11:5,225,465-5,227,070*** (from NCBI)

Gene-Phenotype Relationships

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key
11p15.4	Delta-beta thalassemia	141749	AD	3
	Erythremias, beta-			3
	Heinz body anemias, beta-	140700	AD	3
	Hereditary persistence of fetal hemoglobin	141749	AD	3
	Methemoglobinemias, beta-			3
	Sickle cell anemia	603903	AR	3
	Thalassemia-beta, dominant inclusion-body	603902		3
	Thalassemias, beta-	613985		3
	{Malaria, resistance to}	611162		3

Clinical Synopsis

TEXT

▼ Description

The alpha (HBA1, 141800; HBA2, 141850) and beta (HBB) loci determine the structure of the 2 types of polypeptide chains in adult hemoglobin, HbA. Mutant beta globin that sickles causes sickle cell anemia (603903). Absence of beta chain causes beta-zero-thalassemia. Reduced amounts of detectable beta globin causes beta-plus-thalassemia. For clinical purposes, beta-thalassemia (613985) is divided into thalassemia major (transfusion dependent), thalassemia intermedia (of intermediate severity), and thalassemia minor (asymptomatic).

▼ External Links
▶ Genome
▶ DNA
▶ Protein
▶ Gene Info
▼ Clinical Resources
Clinical Trials
Gene Tests
▶ Genetics Home Reference
GTR
Newborn Screening
GARD
▶ Variation
▶ Animal Models
▶ Cell Lines
▶ Cellular Pathways

https://www.omim.org/entry/141900?search=hemoglobin&highlight=hemoglobin%20haemoglobin

Functions of human genes

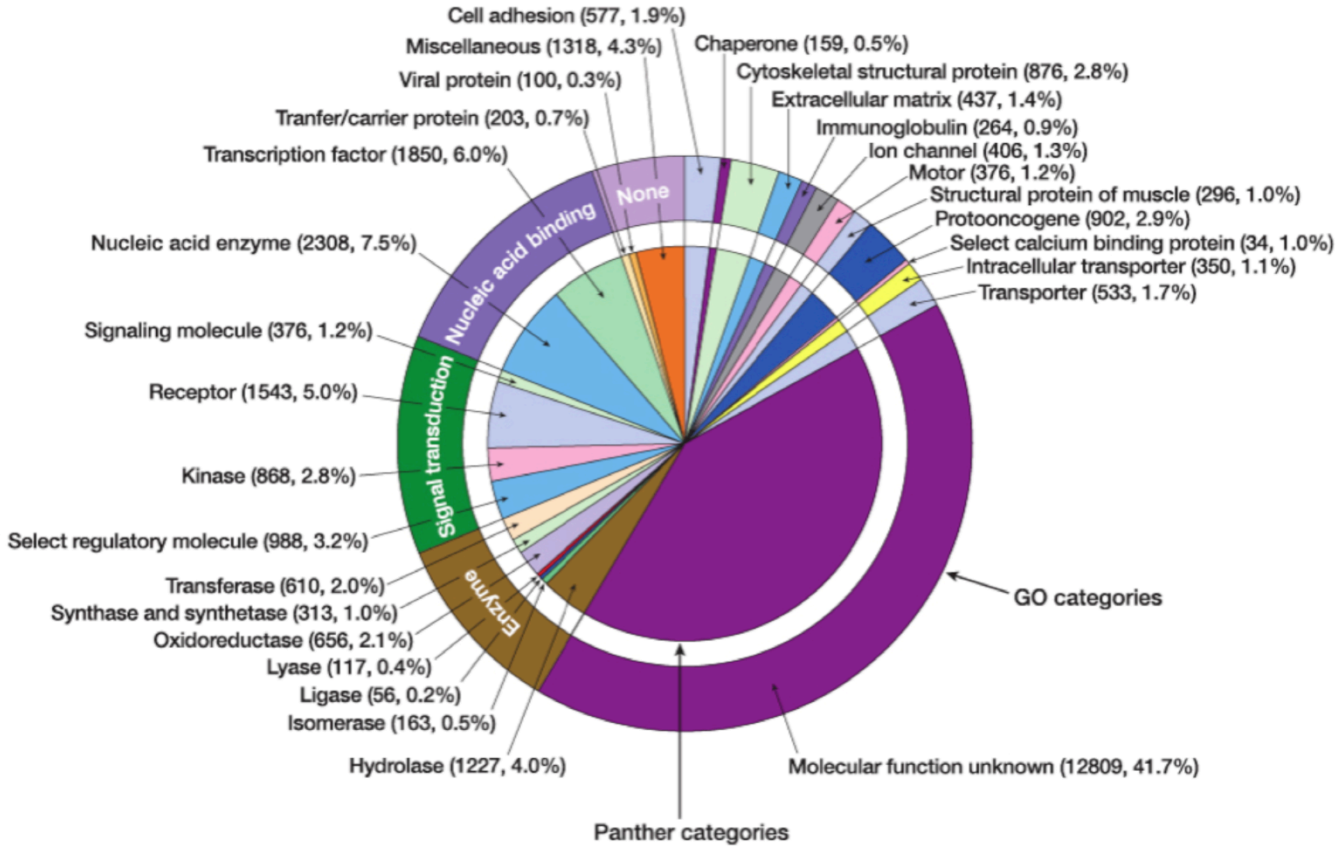


Figure 12.20: A preliminary functional classification of human polypeptide-encoding genes.

Known or predicted functions for 26 383 human polypeptide-encoding genes. Classification is according to the GO molecular function categories as shown in the outer circle (Gene Ontology classification – see Section 8.3.6) or to Celera's Panther molecular function categories (inner circle). Reproduced from Venter *et al.* (2001) *Science* **291**, 1304–1351, with permission from the American Association for the Advancement of Science.

Nomenklatur bei Genveränderungen

SNV

Single Nucleotide ***Variant***

Persönliche Variation des Genoms

SNP

Single Nucleotide ***Polymorphism***

SNV, der in mindestens 1% der untersuchten Population gefunden werden konnte

CNV

Copy number variation

SNVs
Single Nucleotide Variants

..ACGGC..	..ACTGC..
..TGCCG..	..TGACG..

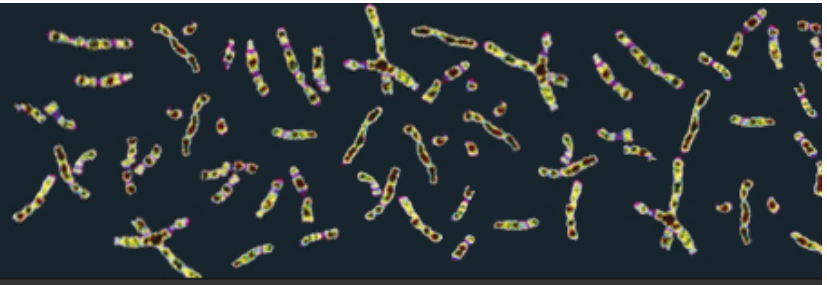
Zwei verglichene Genome unterscheiden sich im Mittel alle **1000 Bp.**

Bei Vergleich aller Genome weltweit schätzt man **9-12 Mio. SNVs (also 1/300 Bp).**

Aus diesen **(und weiteren) Unterschieden** resultiert unsere **Individualität**

1000 Genomes

A Deep Catalog of Human Genetic Variation



LATEST ANNOUNCEMENTS

WEDNESDAY SEPTEMBER 30, 2015

A global reference for human genetic variation

The Phase 3 publication, A global reference for human genetic variation and the Phase 3 Structural variation publication, An integrated map of structural variation in 2,504 human genomes are now available from Nature alongside a celebration of 25 years of the Human Genome Project

The variants from the Phase 3 analysis are available in ftp/release/20130502/ and extended information about the SV dataset can be found in ftp/phase3/integrated_sv_map/.

Both these papers are open access and should be free for everyone to read and download.

If you have any questions about the data these papers are based on or how to access it please email info@1000genomes.org

http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/

Recent project announcements

FRIDAY NOVEMBER 27, 2015

Changes to the 1000 Genomes Globus endpoint

EMBL-EBI has recently rearranged its Globus hosted endpoints.

NAVIGATION

- Frequently Asked Questions

LINKS

All Project Announcements

Sample and Project Information

Media Archive

Find the 1000 Genomes Project Publications

SNPs, Haplotypen und Erkrankungen

ACHTUNG: nur selten ist ein SNP die direkte genetische Ursache für eine Erkrankung oder Prädisposition!

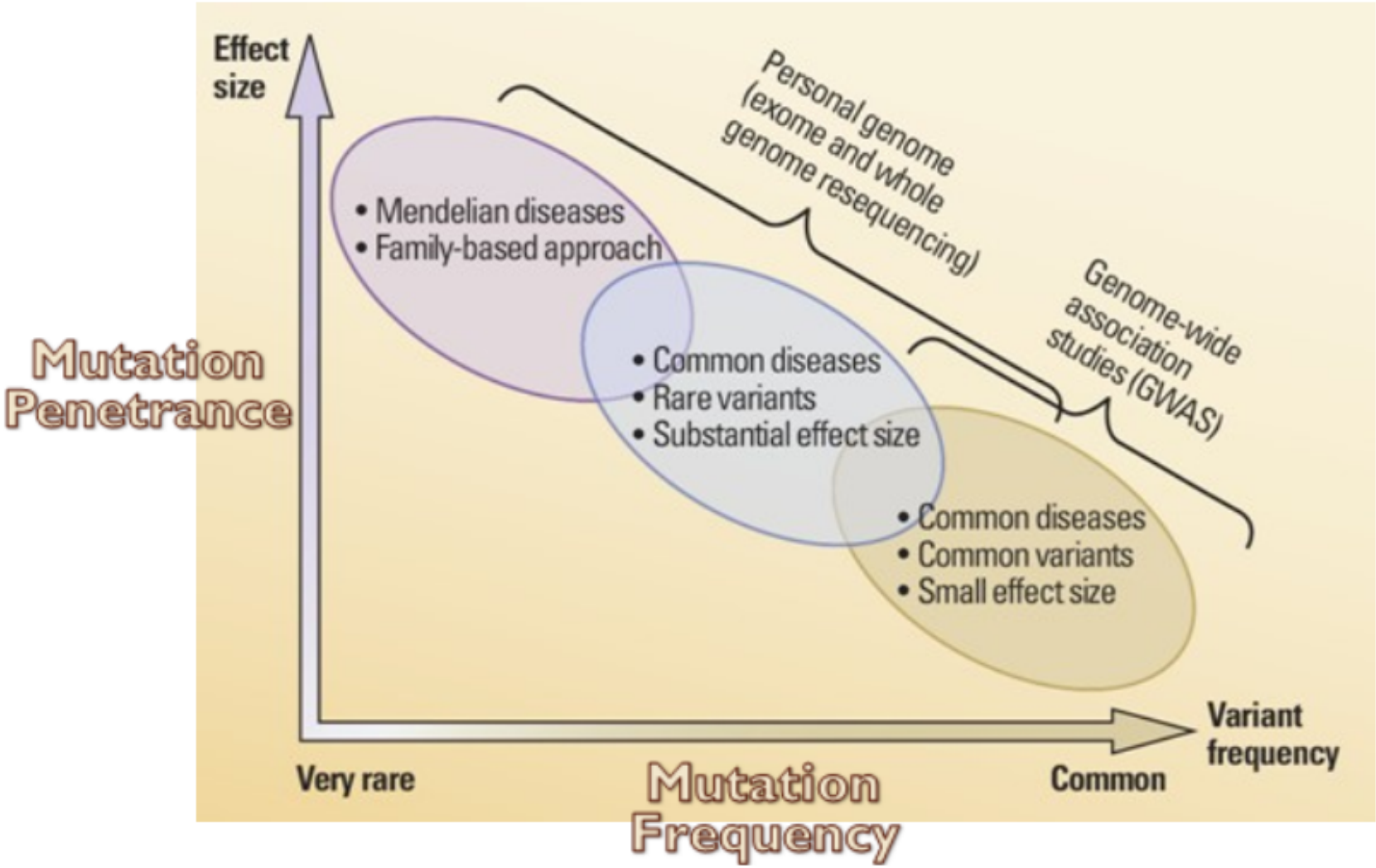
SNPs sind vielmehr als genetische Marker zu sehen, die einen „Haplotyp“ definieren und mit einem unbekannten Krankheitsallel gekoppelt vererbt werden

Genetische Krankheiten

- **Mendel‘sche Krankheiten**
  - (meist) monogen
  - Hohe Penetranz
  - Beispiele: Sichelzellenanämie (HBB), zystische Fibrose (CFTR), Huntington

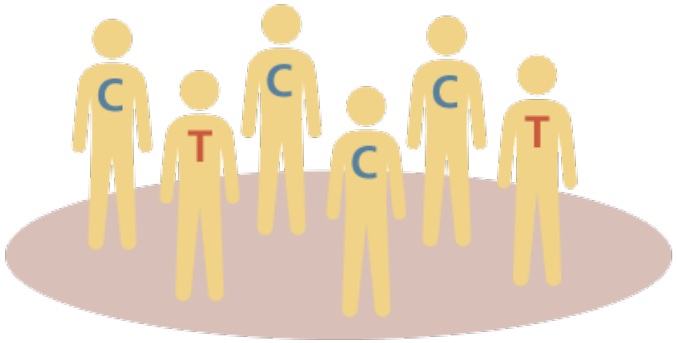
- **Komplexe Krankheiten**
  - Polygen, d.h. durch mehrere Gene bedingt
  - Niedrige Penetranz
  - Multifaktoriell: können durch Umwelt/Ernährung beeinflusst werden
  - Beispiele: Herzkreislauferkrankungen, Diabetes, Übergewicht, Demenz, Krebs, Skelett

SNPs und Erkrankungen

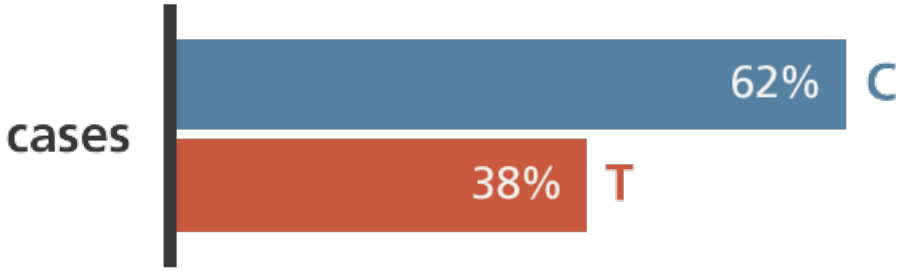


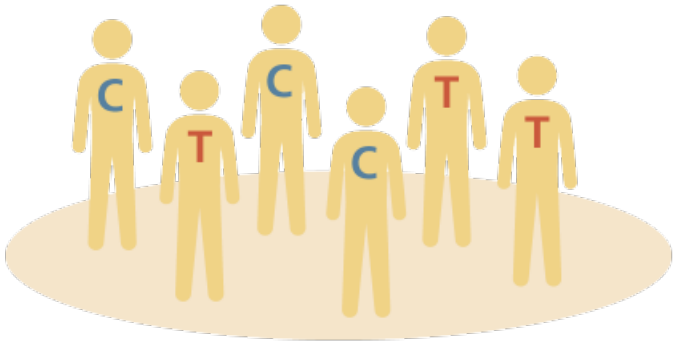
Kaiser J. Science (2012) 338:1016-1017

Genome-wide association studies (GWAS)



cases (n=1,000)
people with heart disease



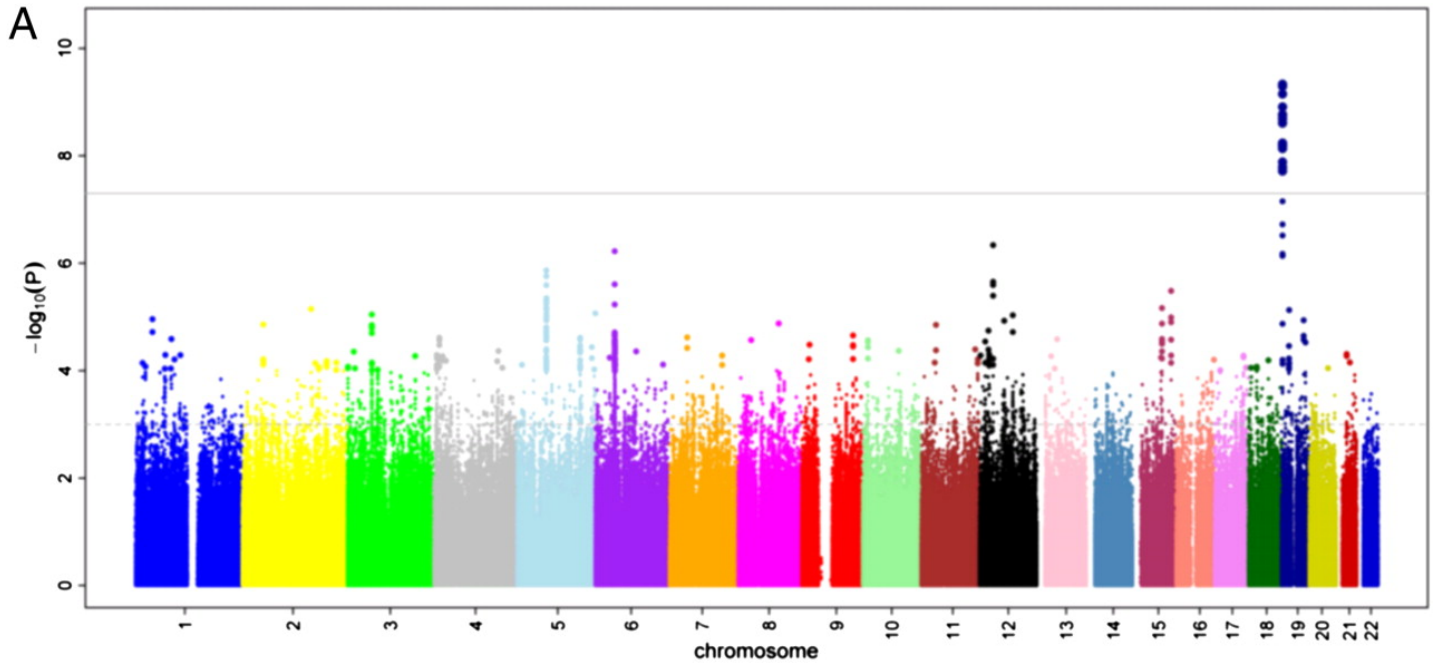


controls (n=1,000)
people without heart disease

http://www.yourgenome.org/stories/genome-wide-association-studies

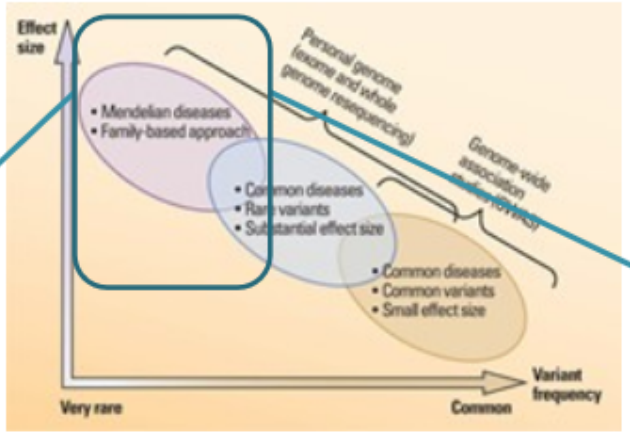
Genome-wide association studies (GWAS)

Genome-wide association and functional studies identify the *DOT1L* gene to be involved in cartilage thickness and hip osteoarthritis



Martha C. Castaño Betancourt et al. PNAS 2012;109:8218-8223

Mendel‘sche Krankheiten

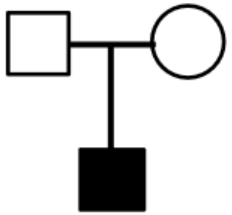


Frequency of disorder		
Rare (<1/10,000)	Low frequency (1/10,000–1/100)	Common (>1/100)
Mutational target		
e.g. CHARGE syndrome (1/10,000)	e.g. Noonan syndrome (1/2,000)	e.g. intellectual disability (2/100)
CHD7	*PTPN11* *RAF1* *SOS1* *KRAS* *BRAF* *MAP2K1* *NRAS*	
Single gene	2–100 genes	>100 genes

Veltman J.A. Nat. Rev. Genetics (2012) 13:565:575.

Mendel‘sche Krankheiten: Gen-Identifizierung durch Trio-Analyse

- Ganze Familie wird sequenziert (gesunde Eltern sowie betroffenes Kind)



- Wieso alle drei?
  - Alle Vererbungsmodelle können untersucht werden
  - Im Gegensatz zur GWAS nur ein einziger Fall nötig!
  - Probenmaterial ist zugänglich

Exom-Sequenzierung !!

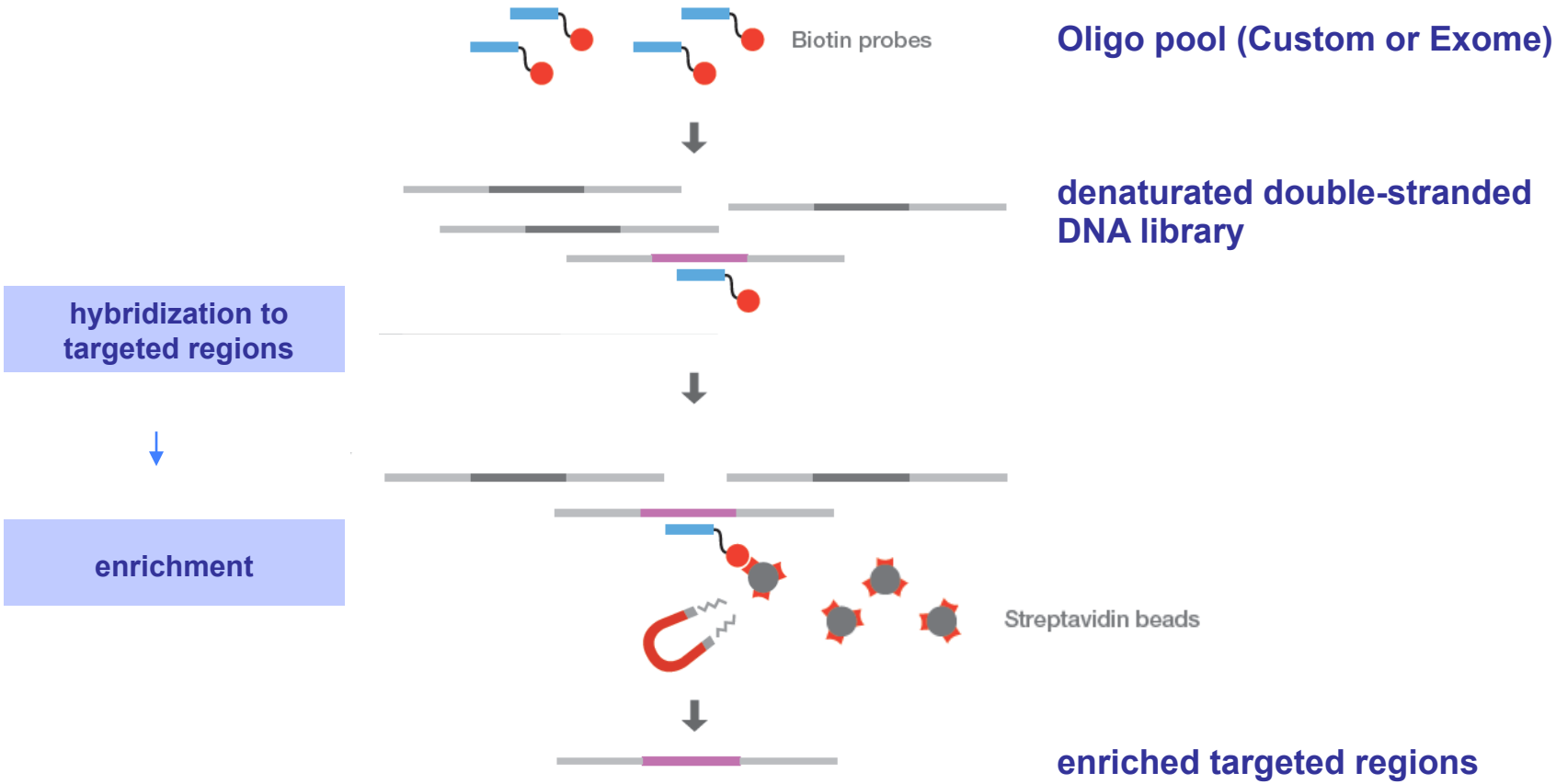
- 3.000.000.000 bp (3 Gb) Humangenom
  - Ca. 45 % repetitiv
  - Ca. 25 % Genregionen
  - Ca. 2% Exons bzw. codierende Regionen

- 20.000 – 30.000 humane Gene
  - 3000-5000 sind Krankheitsgene
  - Ca. 4000 humane genetische Krankheiten
  - 114 Gene, deren Fehlfunktion tatsächlich mit Medikamenten behandelt werden kann

Michael O. Dorschner, Am J Hum Genet (2013) 93:631-640

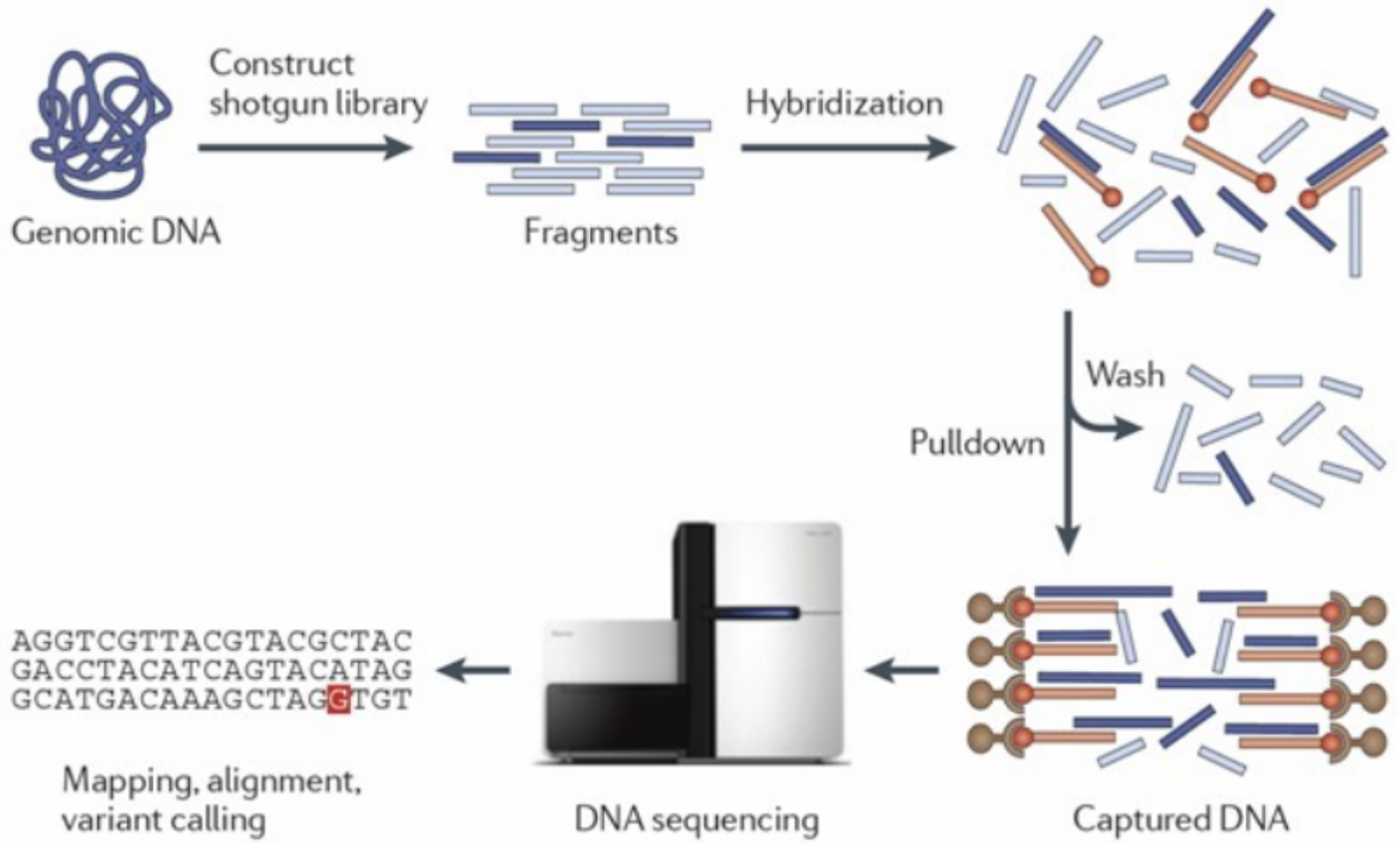
Targeted Resequencing:

Custom/ Exome Enrichment



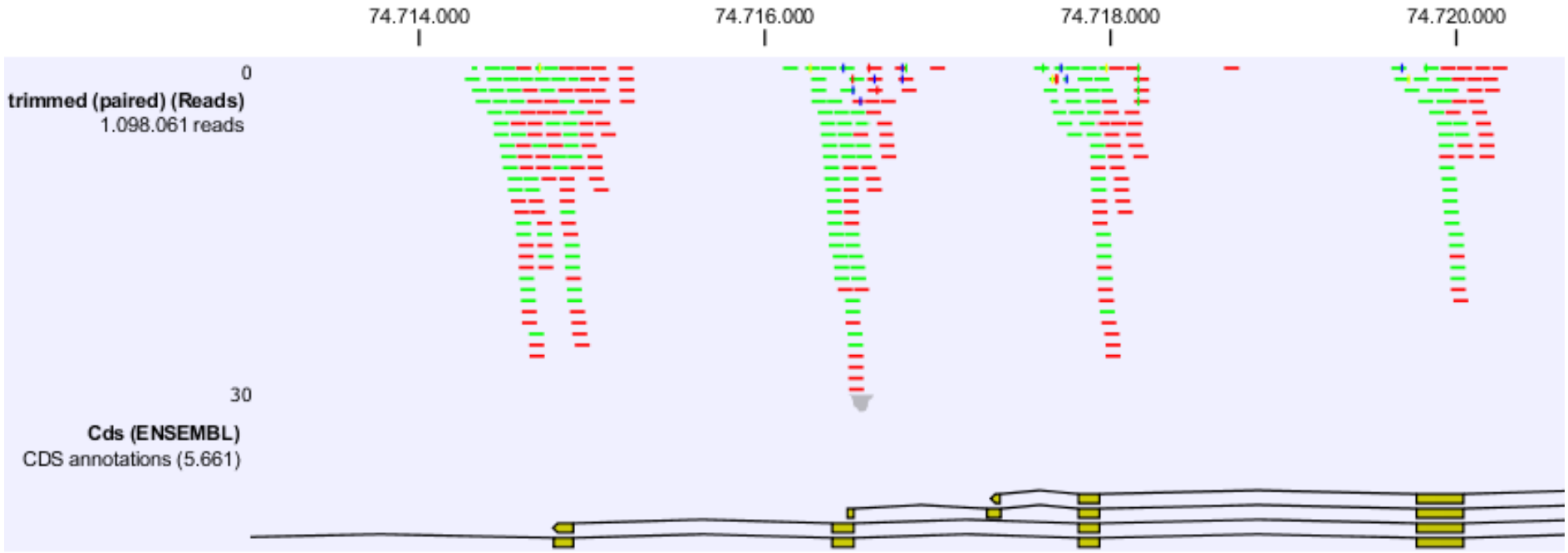
B. Ebner

Targeted Resequencing



Bamshad MJ., Nat. Rev. Genetic (2011) 12:745-755

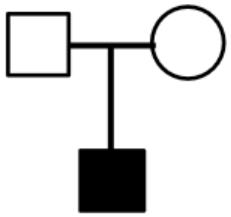
Targeted Resequencing: Mapping



B. Ebner

Exomsequenzierung eines „Trios“

- Ganze Familie wird sequenziert (gesunde Eltern so wie betroffenes Kind)



- Wieso alle 3?
  - Alle Vererbungsmodi können untersucht werden
  - Im Gegensatz zur GWAS nur ein einziger Fall nötig!
  - Probenmaterial ist zugänglich

Die Idee dahinter?

- Identifikation von SNVs, die spezifisch für die Familie sind

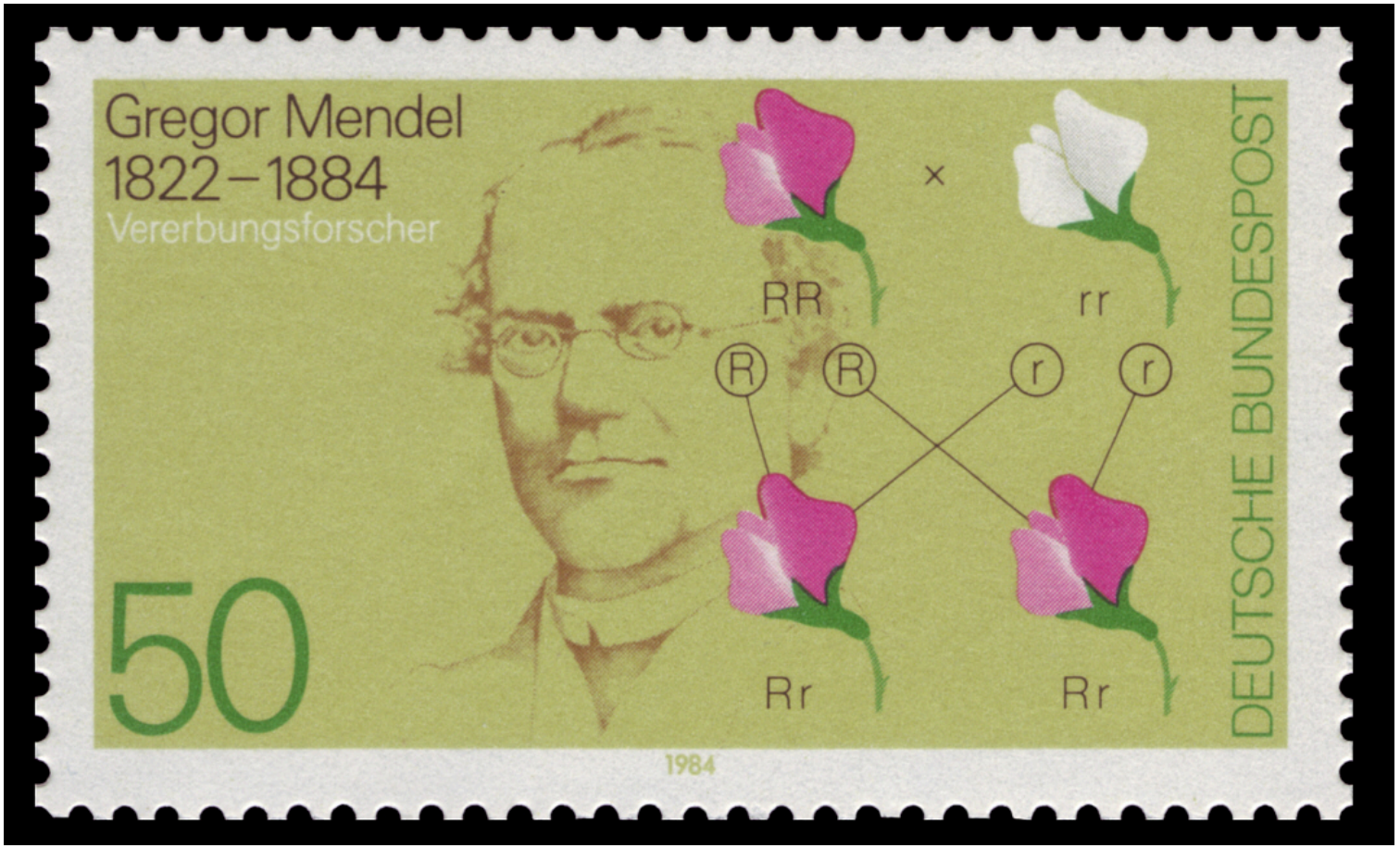
- Suche nach dem „Schuldigen“ durch das Ausschlussprinzip:

    - Welche Mutationen sind neu und potenziell gefährlich?
    - An welchen Stellen wurden z.B. Varianten an das Kind homozygot weitergegeben, für die die Eltern heterozygot waren?

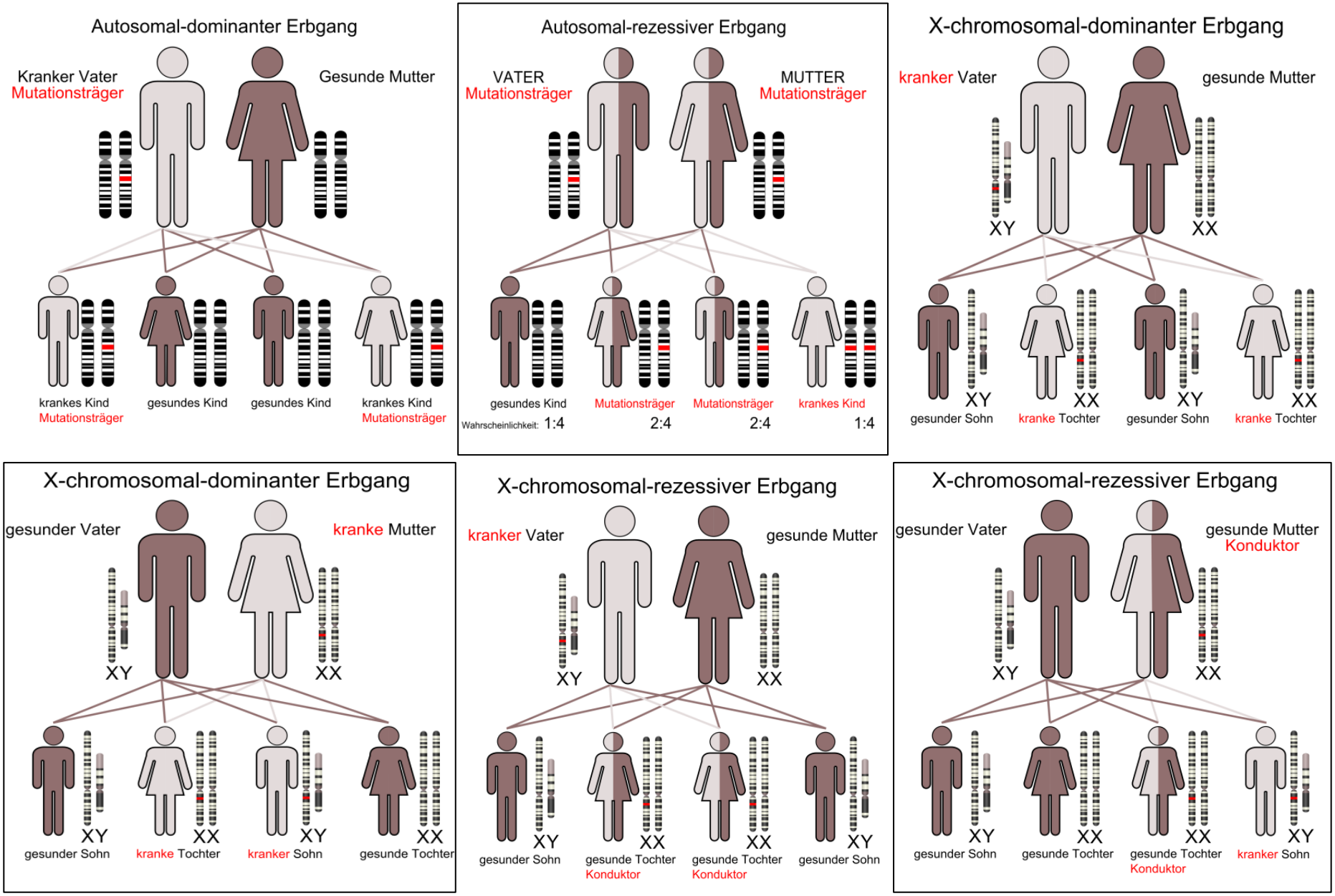
- Falls Parental- und Filialgeneration betroffen sind: Welche Varianten sind vom kranken Elternteil vererbt und kommen im gesunden nicht vor?

- Optimal: zusätzliche Sequenzierung von Geschwistern

Vererbungsmodi? Mendel!!



Entdecke die Möglichkeiten...

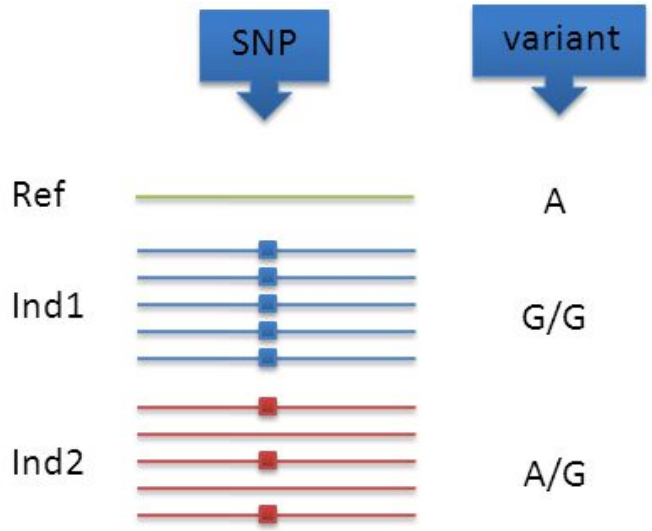


www.wikipedia.org

Diagnostische Detektion von SNVs: technischer Ablauf

1) Read mapping: Kartierung der Exom-Reads gegen das humane Referenzgenom

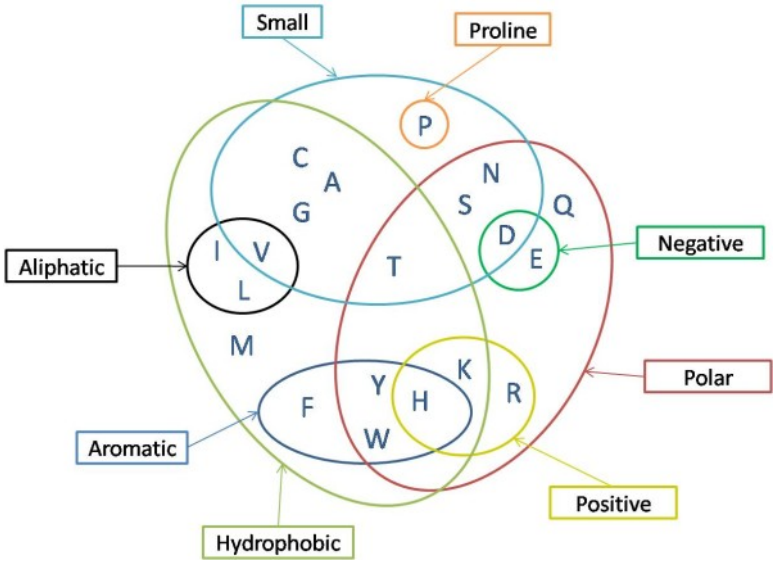
2) Identifikation der Varianten:

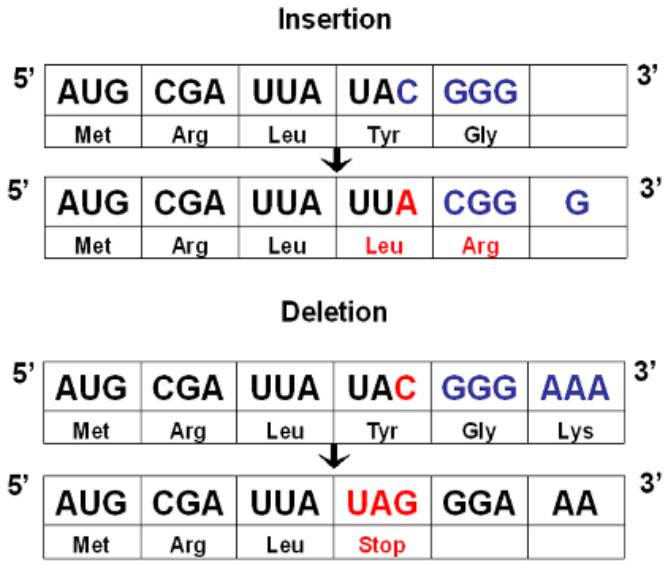


Diagnostische Detektion von SNVs: technischer Ablauf

3) Annotation der Varianten:

→ Detektion von veränderten Aminosäuresequenzen
→ Abgleich mit Informationen aus öffentlichen Datenbanken

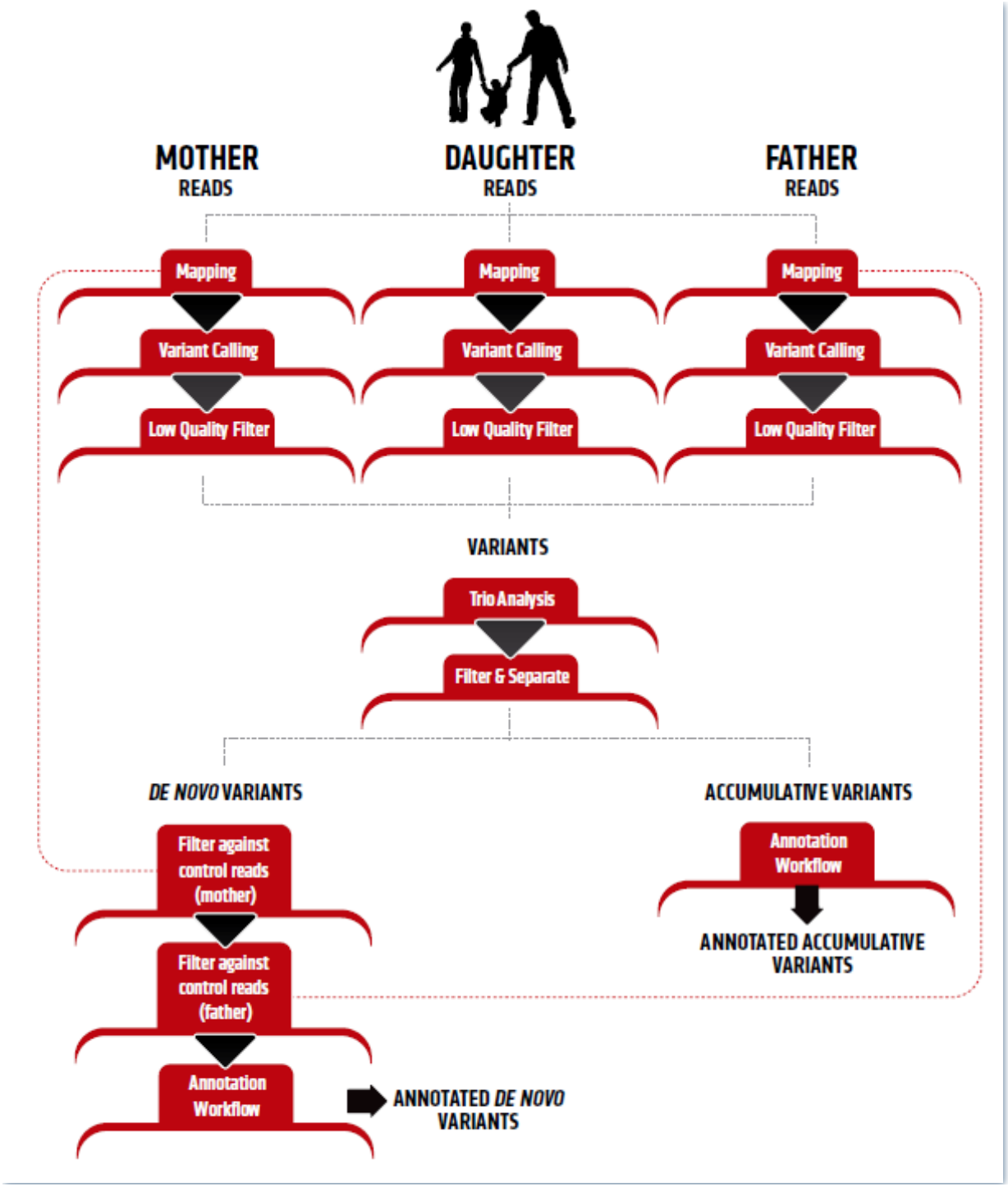




Exarchos, Konstantinos et al (2009) BMC bioinformatics. 10. 113. DOI: 10.1186/1471-2105-10-113.

https://sbi4u2013.wordpress.com/2013/02/10/gene-muterations/

Technischer Ablauf



Kursszenario

Familie A

- 2 Kinder (2 Mädchen)
- Eines der Kinder zeigt Auffälligkeiten im Verhalten und der Entwicklung

Familie B

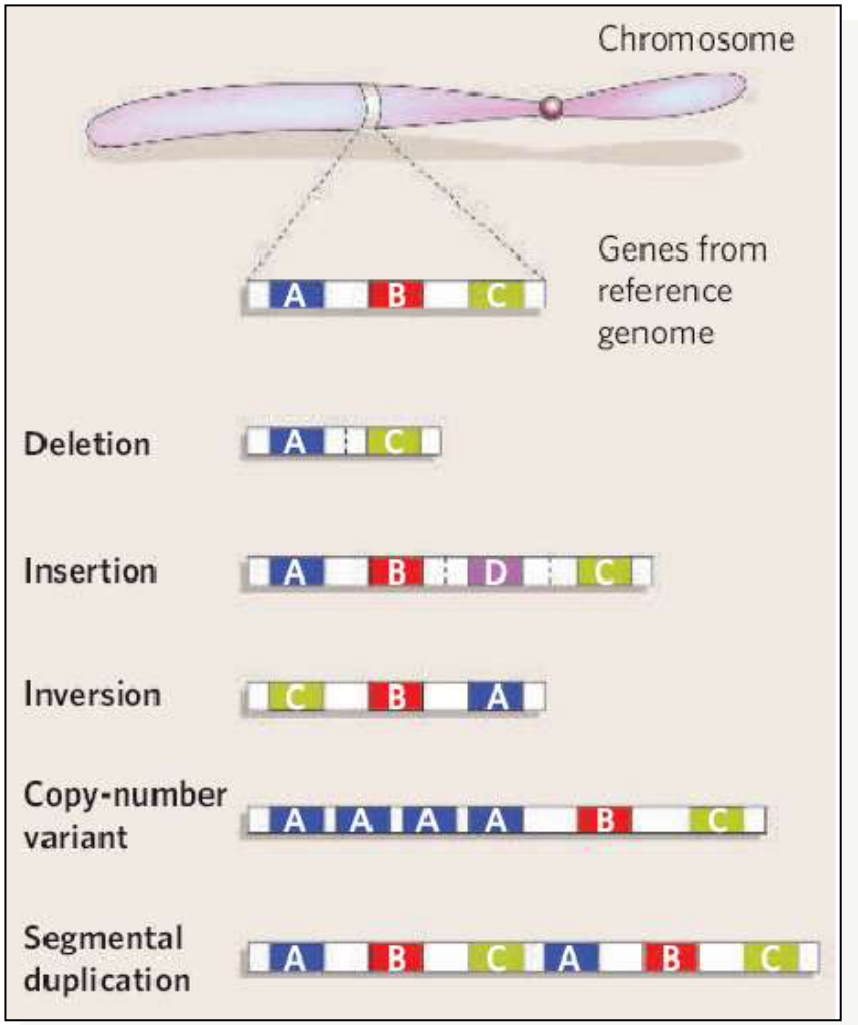
- 2 Kinder (Mädchen und Junge)
- Junge zeigt Auffälligkeiten im Verhalten und der Entwicklung → große Ähnlichkeit zur ersten Familie!

Keine Verwandtschaft zwischen den Eltern beider Familien, keine bekannte, krankheitsrelevante Vorgeschichte in beiden Fällen!

Kursszenario

Familie 11114 (A)		**Familie 11892 (B)**	
SRR1272259	Vater	SRR1301491	Vater
SRR1272260	Mutter	SRR1301492	Mutter
SRR1272261	Krankes Kind (weibl.)	SRR1301493	Krankes Kind (männl.)
SRR1272262	Gesundes Kind (weibl.)	SRR1301494	Gesundes Kind (weibl.)

Aber: SNPs/SNVs sind nicht alles!



Copy number variations

	CNV (Database of Genomic Variants, http://projects.tcag.ca/variation/)	SNP (dbSNP, http://www.ncbi.nlm.nih.gov/SNP/)
Total number	38,406[a] (Mar 11, 2009)	14,708,752 (Build 129)
Size	100 bp to 3 Mb	Mostly 1 bp
Type	Deletion, duplication, complex	Transition, transversion, short deletion, short insertion
Effects on genes	Gene dosage, interruption, etc.	Missense, nonsense, frameshift, splice site
Percentage of the reference genome covered	29.74%[b]	<1%

Mutation rates for CNVs are locus-specific
and can be 100 – 10 000 fold higher than for SNPs

Zhang et al. (Ann Rev Genom Hum Gen 2009)

Structural variation & disease

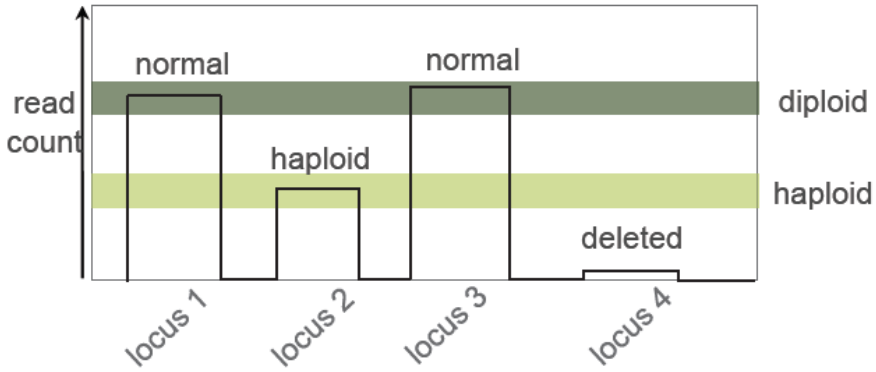
Phenotype	OMIM	Locus	CNV
Mendelian (autosomal dominant)[b]			
Williams-Beuren syndrome	194050	7q11.23	del
7q11.23 duplication syndrome	609757	7q11.23	dup
Spinocerebellar ataxia type 20	608687	11q12	dup
Smith-Magenis syndrome	182290	17p11.2/*RAI1*	del
Potocki-Lupski syndrome	610883	17p11.2	dup
HNPP	162500	17p12/*PMP22*	del
CMT1A	118220	17p12/*PMP22*	dup
Miller-Dieker lissencephaly syndrome	247200	17p13.3/*LIS1*	del
Mental retardation	601545	17p13.3/*LIS1*	dup
DGS/VCFS	188400/192430	22q11.2/*TBX1*	del
Microduplication 22q11.2	608363	22q11.2	dup
Adult-onset leukodystrophy	169500	*LMNB1*	dup
Mendelian (autosomal recessive)			
Familial juvenile nephronophthisis	256100	2q13/*NPHP1*	del
Gaucher disease	230800	1q21/*GBA*	del
Pituitary dwarfism	262400	17q	
Spinal muscular atrophy	253300	5q1	
beta-thalassemia	141900	11p	
alpha-thalassemia	141750	16p	
Complex traits			
Alzheimer disease	104300	*APP*	dup
Autism	612200	3q24	inherited homozygous del
	611913	16p11.2	del/dup
Crohn disease	266600	*HBD-2*	copy number loss
	612278	*IRGM*	del
HIV susceptibility	609423	*CCL3L1*	copy number loss
Mental retardation	612001	15q13.3	del
	610443	17q21.31	del
	300534	Xp11.22	dup
Pancreatitis	167800	*PRSS1*	tri
Parkinson disease	168600	*SNCA*	dup/tri

Need for high-resolution detection of CNVs (and SNVs) by DNA sequencing

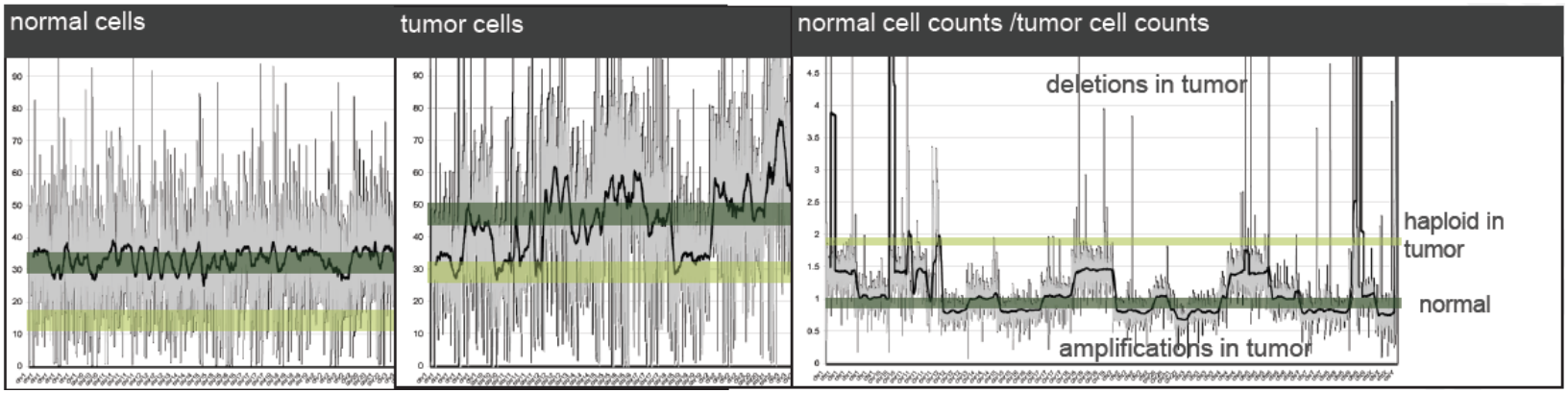
Zhang et al. (Ann Rev Genom Hum Gen 2009)

Detektion von CNVs

copy number variations



Vergleich der CNV-Profile gesunder Zellen und Krebszellen eines Patienten:



Detektion von Genom-Rearrangements

mapping

of paired-end reads against genome

→

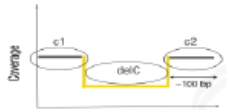
filtering

for read pairs with distance outside expected range

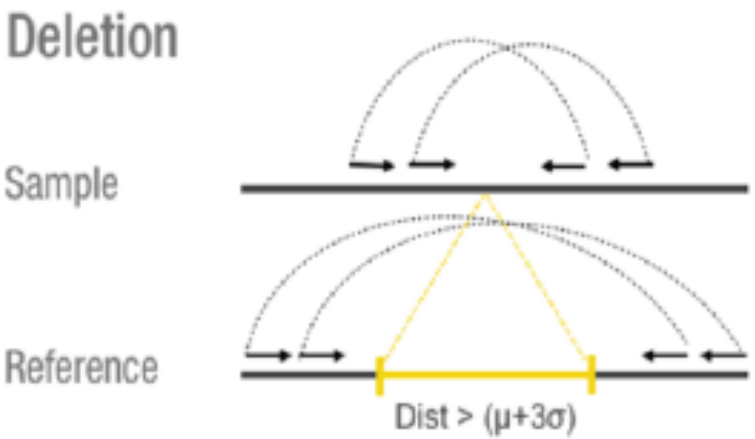
Look for support →

Gapped alignment

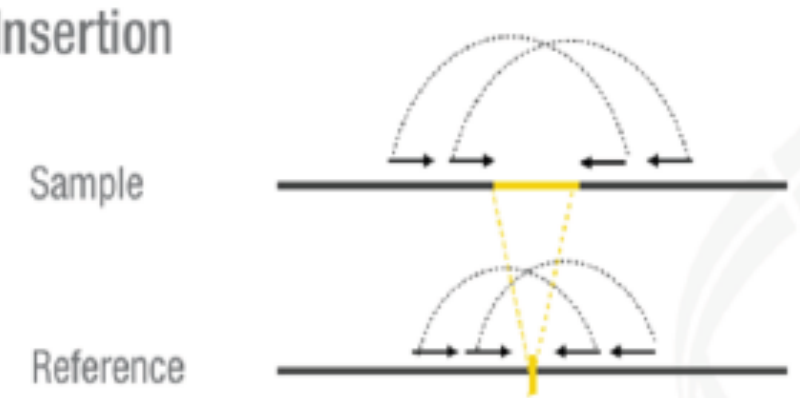
Density plots



Deletion



Insertion



Detektion von Genom-Rearrangements

mapping

of paired-end reads against genome

→

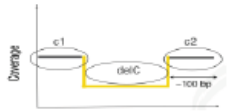
filtering

for read pairs with distance outside expected range

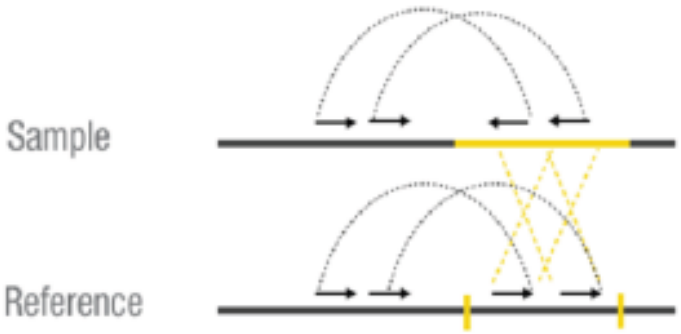
Look for support →

Gapped alignment

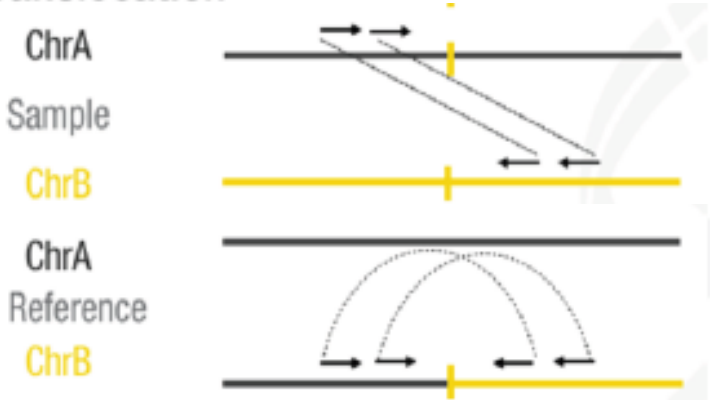
Density plots



Inversion



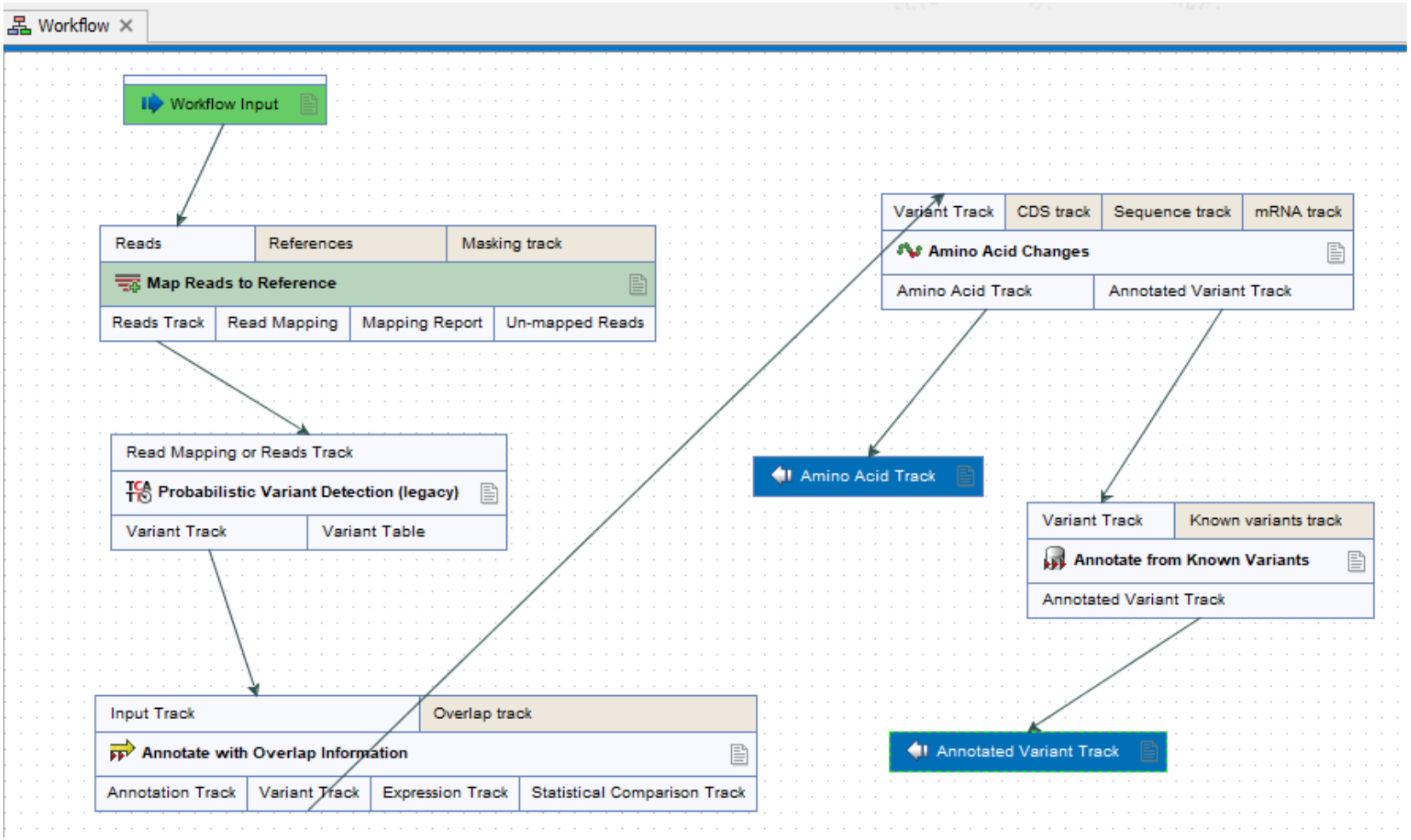
Translocation



Trioanalyse - Praxisteil

Screenshots und Einstellungen

Der CLC-Workflow



CLC-Export:
Diese Spalten müssen dabei sein!

Position

gene_name (Homo sapiens (hg19)_Gene)

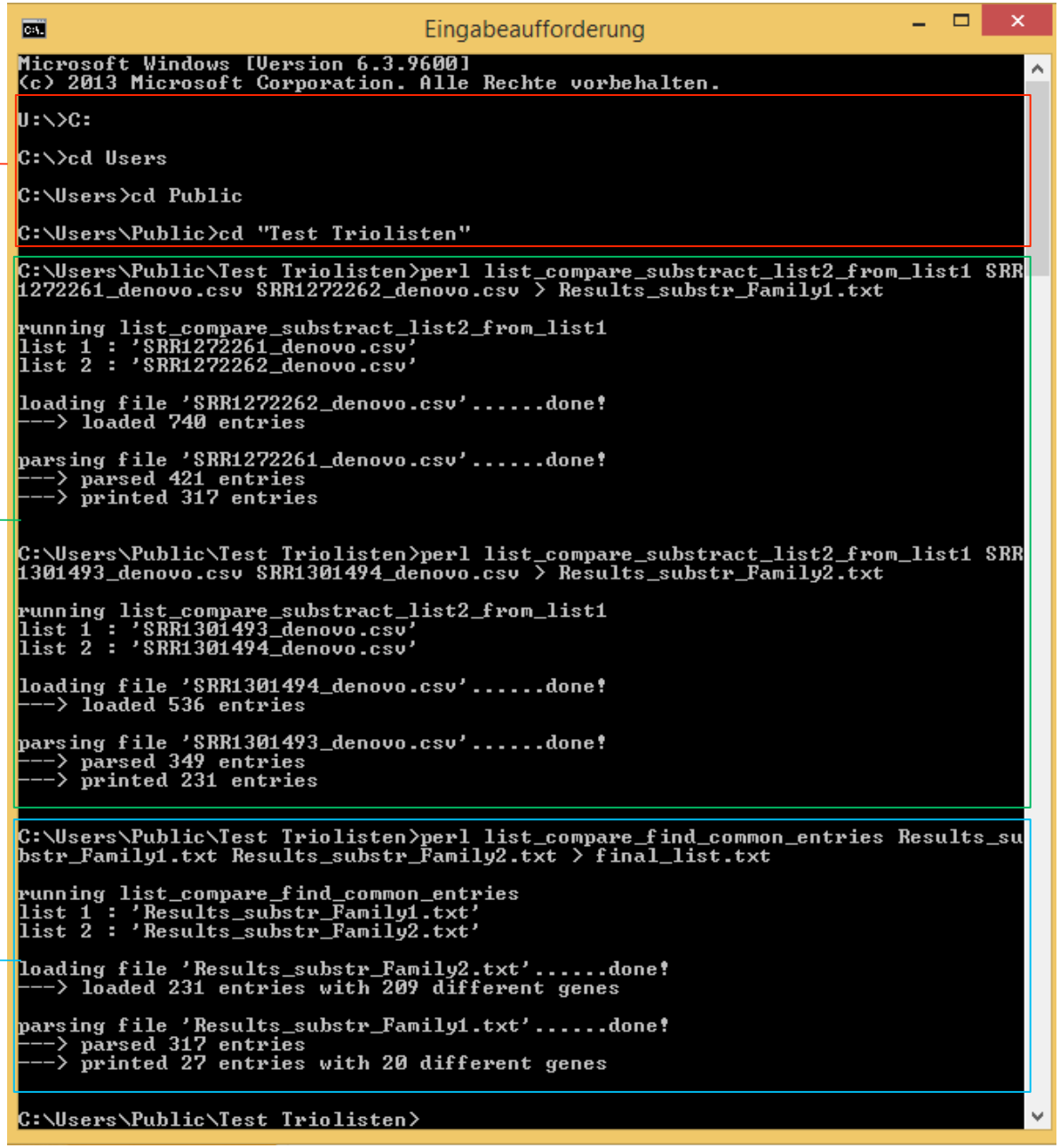
Amino acid change in longest transcript

Coding region change in longest transcript

Navigiere in den Ordner, in dem die Listen hinterlegt sind

Vergleiche die Geschwisterkinder untereinander und suche nach *Unterschieden*

Vergleiche die kranken Kinder von beiden Familien und finde *Gemeinsamkeiten*



Detektion von Mutationen des kranken Kindes, die das gesunde Geschwisterkind nicht hat

```perl
#!/usr/bin/perl

##########  check input  ##########

$0 =~ s/.*\/(.+)/$1/;
unless (@ARGV >= 2)
{
	print STDERR "\n\n$0\n";
	print STDERR "-" x length ($0);
	print STDERR "\nscript to compare two tab separated lists, substract list2 & prints entries from list 1 when conform with column 1 & 2\n";
	print STDERR "results given on STDOUT (>)\n";
	print STDERR "\n";
	print STDERR "USAGE\n";
	print STDERR "$0 [list1] [list2]\n";
	print STDERR "\n\n";
	exit;
}

my $list1 = $ARGV[0];
my $list2 = $ARGV[1];

die "\n$list1: no such file!\n\n" unless (-f $list1);
die "\n$list2: no such file!\n\n" unless (-f $list2);

##########  /check input  ##########

print STDERR "\nrunning $0\n";
print STDERR "list 1 : '$list1'\n";
print STDERR "list 2 : '$list2'\n";

my %LIST1;
my $list1_entry_sum = 0;
my $list1_printed = 0;
my %LIST2;
my $list2_entry_sum = 0;

print STDERR "\nloading file '$list2'...";

open (DAT, "<", "$list2") || die "\ncouldn't open file '$list2'!\n\n";

while (<DAT>)
{
	chomp ($_);
	my @L = split (/;/, $_);

	#####  WORK IN HERE  #####

	$LIST2{$L[1]}{$L[0]}++;
	$list2_entry_sum++;

	#####  /WORK IN HERE  #####
}

close (DAT);

print STDERR "\rloading file '$list2'......done!\n";
print STDERR "---> loaded $list2_entry_sum entries\n";

print STDERR "\nparsing file '$list1'...";

my %GENES_PRINTED_SUM;

open (DAT, "<", "$list1") || die "\ncouldn't open file '$list1'!\n\n";

while (<DAT>)
{
	chomp ($_);
	my @L = split (/;/, $_);

	#####  WORK IN HERE  #####

	$list1_entry_sum++;

	next if (exists ($LIST2{$L[1]}) && exists ($LIST2{$L[1]}{$L[0]}));

	print "$_\n";
	$list1_printed++;
	$GENES_PRINTED_SUM{$L[1]}++;

	#####  /WORK IN HERE  #####
}

close (DAT);

print STDERR "\rparsing file '$list1'......done!\n";
print STDERR "---> parsed $list1_entry_sum entries\n";
print STDERR "---> printed $list1_printed entries\n\n";
```

Schnittmenge der mutierten Gene beider kranken Kinder

```perl
#!/usr/bin/perl

##########  check input  ##########

$0 =~ s/.*\/(.+)/$1/;
unless (@ARGV >= 2)
{
	print STDERR "\n\n$0\n";
	print STDERR "-" x length ($0);
	print STDERR "\nscript to compare two tab separated lists and print entries from list 1 sharing common column (2) with list 2\n";
	print STDERR "results given on STDOUT (>)\n";
	print STDERR "\n";
	print STDERR "USAGE\n";
	print STDERR "$0 [list1] [list2]\n";
	print STDERR "\n\n";
	exit;
}

my $list1 = $ARGV[0];
my $list2 = $ARGV[1];

die "\n$list1: no such file!\n\n" unless (-f $list1);
die "\n$list2: no such file!\n\n" unless (-f $list2);

##########  /check input  ##########

print STDERR "\nrunning $0\n";
print STDERR "list 1 : '$list1'\n";
print STDERR "list 2 : '$list2'\n";

my %LIST1;
my $list1_entry_sum = 0;
my $list1_printed = 0;
my %LIST2;
my $list2_entry_sum = 0;

print STDERR "\nloading file '$list2'...";

open (DAT, "<", "$list2") || die "\ncouldn't open file '$list2'!\n\n";

while (<DAT>)
{
	chomp ($_);
	my @L = split (/;/, $_);

	#####  WORK IN HERE  #####

	$LIST2{$L[1]}{$_}++;
	$list2_entry_sum++;

	#####  /WORK IN HERE  #####
}

close (DAT);

print STDERR "\rloading file '$list2'......done!\n";
print STDERR "---> loaded $list2_entry_sum entries with ".keys (%LIST2)." different genes\n";

print STDERR "\nparsing file '$list1'...";

my %GENES_PRINTED_SUM;

open (DAT, "<", "$list1") || die "\ncouldn't open file '$list1'!\n\n";

while (<DAT>)
{
	chomp ($_);
	my @L = split (/;/, $_);

	#####  WORK IN HERE  #####

	$list1_entry_sum++;

	if (exists ($LIST2{$L[1]}))
	{
		print "$_\n";
		$list1_printed++;
		$GENES_PRINTED_SUM{$L[1]}++;
	}

	#####  /WORK IN HERE  #####
}

close (DAT);

print STDERR "\rparsing file '$list1'......done!\n";
print STDERR "---> parsed $list1_entry_sum entries\n";
print STDERR "---> printed $list1_printed entries with ".keys (%GENES_PRINTED_SUM)." different genes\n\n";
```