Bio 5491 - Advanced Genetics

Human Genetics Lecture #1
March 31, 2016

Cristina de Guzman Strong, Ph.D.
Department of Medicine
Dermatology/Pharmacogenomics
Center for the Study of Itch
McDonnell Basic Sciences 770
cristinastrong@wustl.edu
362-7695

Human Genetics

- Karyotype
- Genetic Variants
- Mendelian Diseases
- Penetrance/Expressivity
- Human Genome Project
- GWAS/Next-Gen sequencing (Exome)
- Undiagnosed Diseases
- Epigenetics/ENCODE
- Mutations in Regulatory Elements
- Copy number variation diseases
- Mitochondrial genetics
- Human-specific variation
- Future

What is Human Genetics?

The relationship between natural DNA sequence variation(s) and human phenotypic traits

What is different about Human Genetics?

- Imprinting……..uniquely mammalian.

- Trinucleotide repeat diseases…….anticipation.

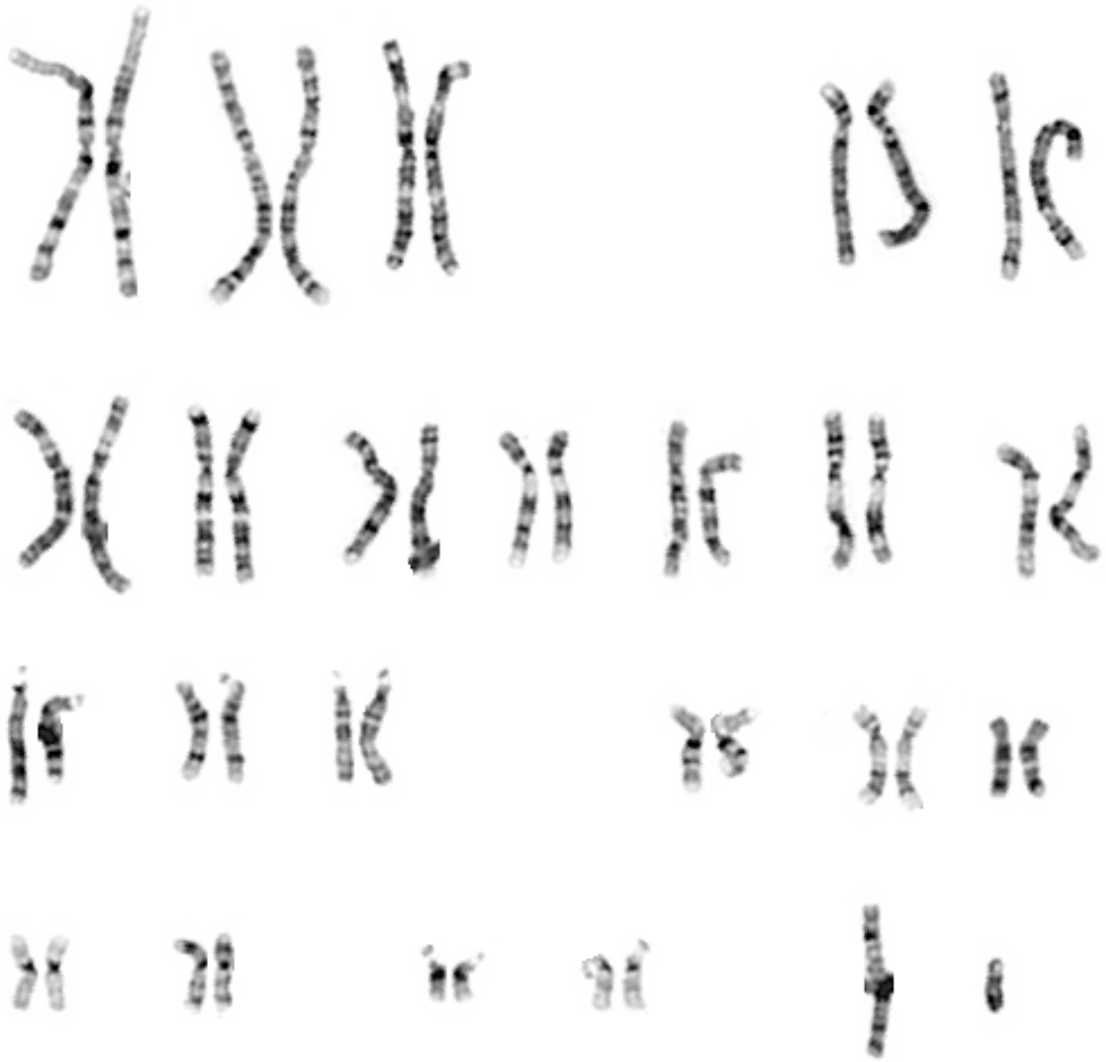
- One can study complex behaviours and cognition.

- Extensive sequence variation leads to common/ complex disease

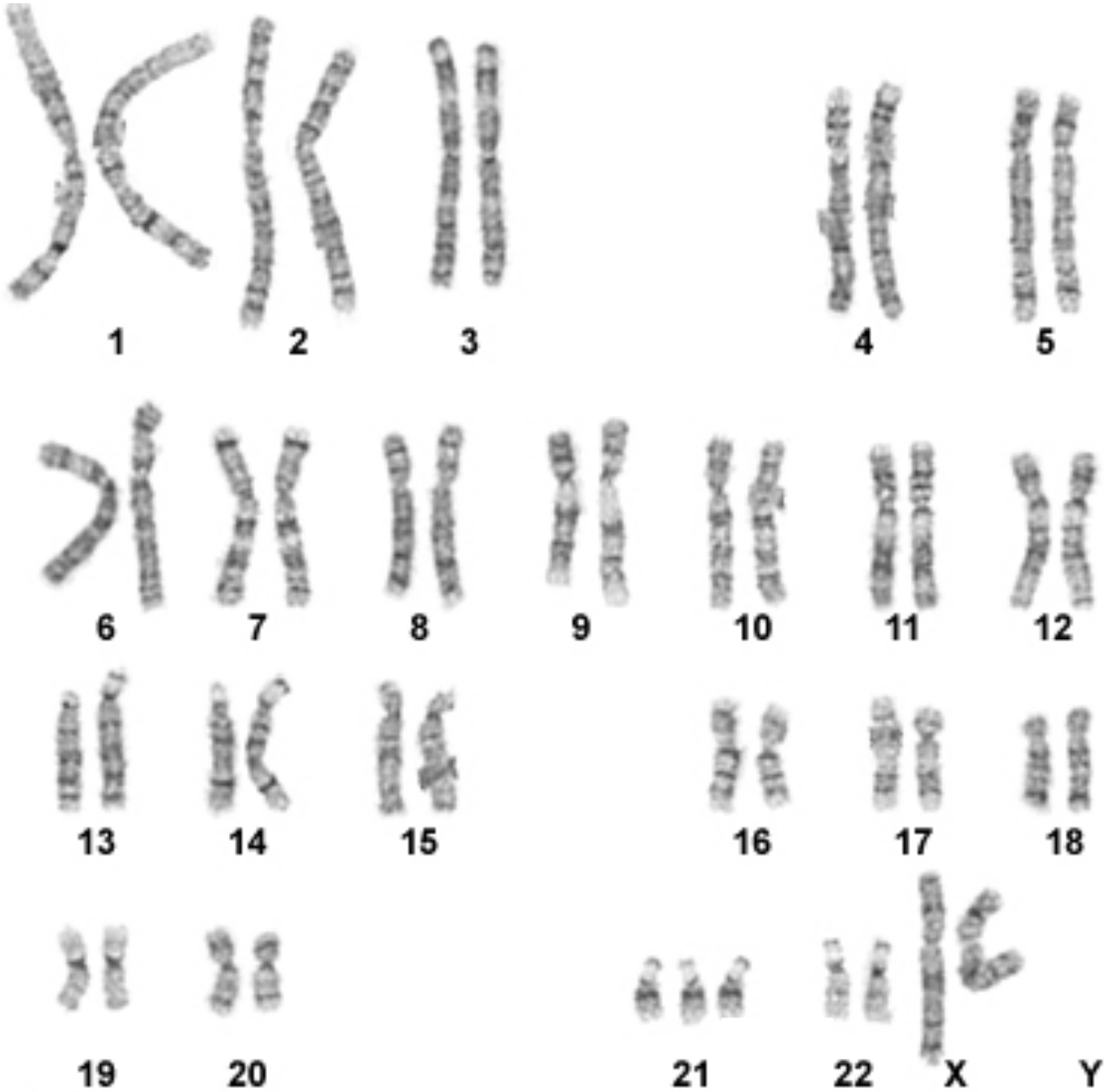
1. Common disease – common variant hypothesis
2. Large # of small-effect variants
3. Large # of large-effect rare variants
4. Combo of genotypic, environmental, epigenetic interactions

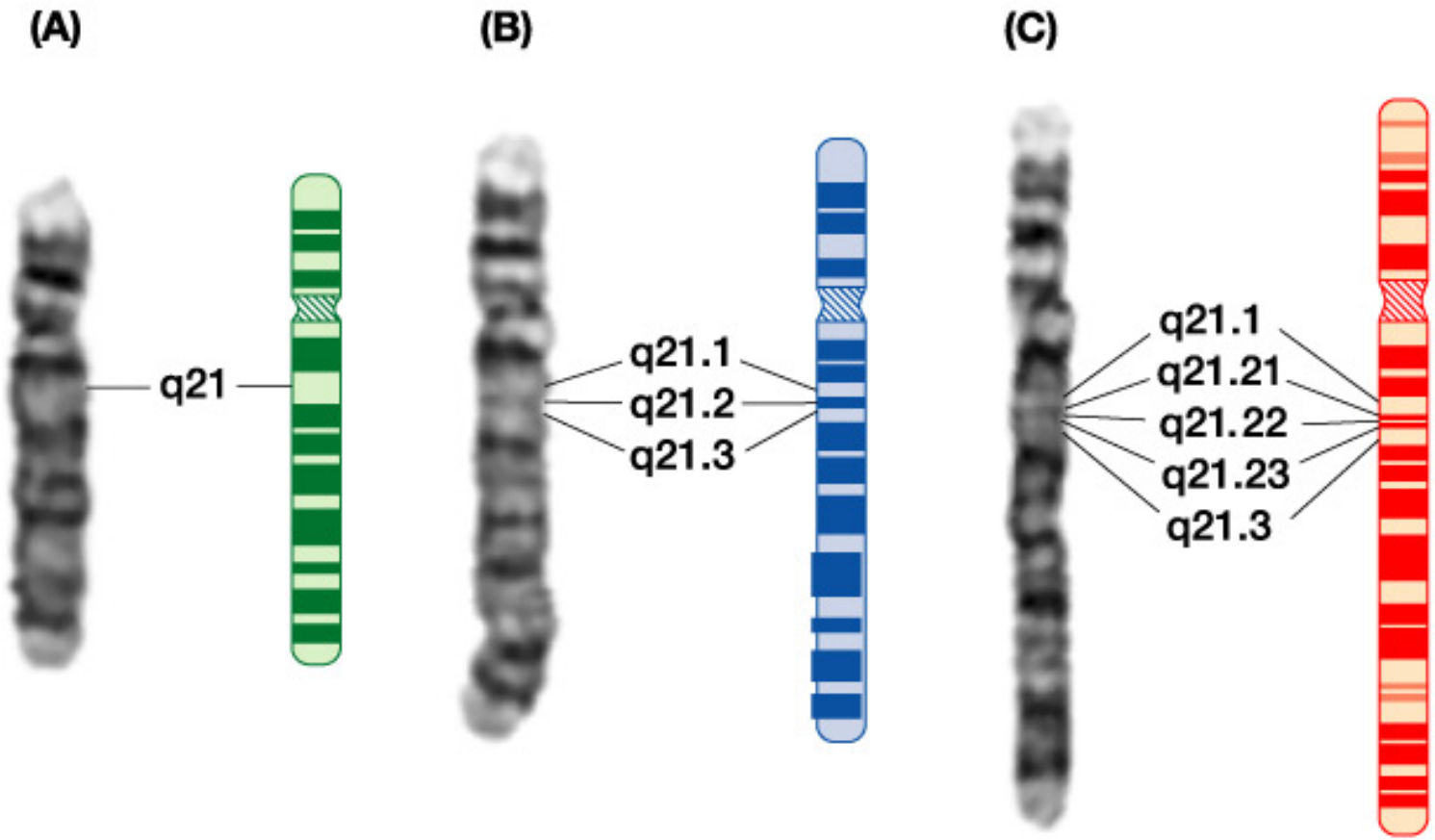
Greg Gibson, Nature Review Gen 2012

Human Genome (Karyotype)



22 autosomes/ XY sex chromosomes

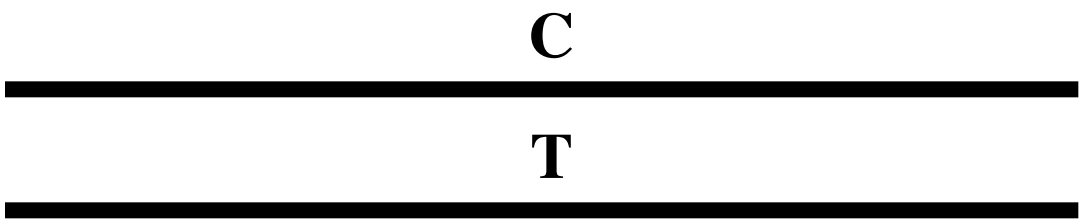




Human genome is ~41% GC, but that is non-randomly distributed. Dark G-bands are lower GC (and lower gene content)

Figure 2-13 Human Molecular Genetics, 3/e. (© Garland Science 2004)

Genetic variation: Single Base Pair



- SNP (single nucleotide polymorphism) - Freq > 0.01
- Can also be 1 insertion or 1 deletion, “indel”
- Alleles with Freq < 0.01 – called rare variants OR SNVs
- Mutations: usually, really RARE. Alter protein function or regulation. Simple Mendelian Trait. Can cause <u>disease</u>

Understanding the SNP
Major vs. Minor Allele

- Referring to coding OR non-coding region
- Nomenclature – rs”X” (SNP), can also be indel

Looking up a SNP

C

T

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data

Simple Nucleotide Polymorphisms (dbSNP build 130)

dbSNP build 130 rs7927894

dbSNP: rs7927894
Position: chr11:75978964-75978964
Band: 11q13.5
Genomic Size: 1
View DNA for this feature (hg18/Human)

Summary: C>C/T (chimp allele displayed first, then '>', then human alleles)
Strand: +
Observed: C/T
Reference allele: C

Chimp allele:	C	**Chimp strand:**	+	**Chimp position:**	chr11:74979721-74979721
Orangutan allele:	C	**Orangutan strand:**	+	**Orangutan position:**	chr11:72059422-72059422
Macaque allele:	C	**Macaque strand:**	+	**Macaque position:**	chr14:74763970-74763970

Looking up a SNP

C

T

e!Ensembl east

BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

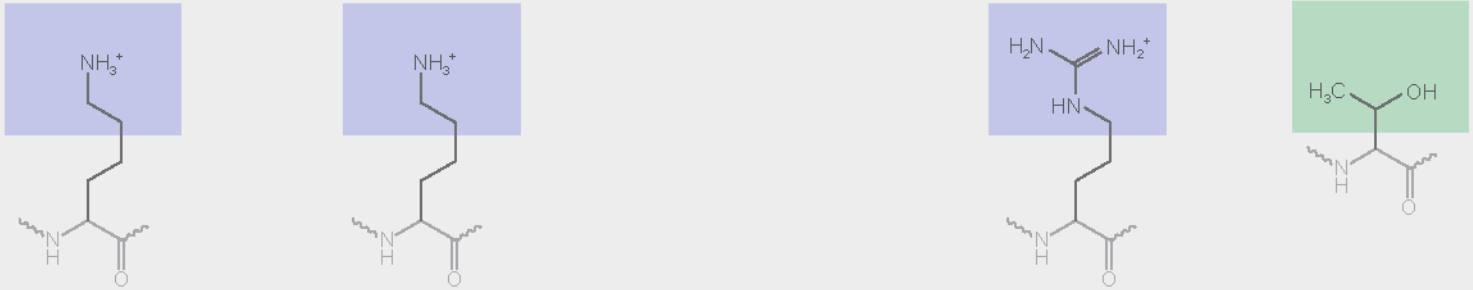
Human (GRCh37) ▼ | Location: 11:76,301,266-76,301,366 | Variation: rs7927894

rs7927894 SNP

Original source	Variants (including SNPs and indels) imported from dbSNP (release 138) | View in dbSNP
Alleles	C/T | Ancestral: **C** | Ambiguity code: **Y** | MAF: **0.28** (T)
Location	Chromosome **11:76301316** (forward strand) | View in location tab
Most severe consequence	Upstream gene variant | See all predicted consequences [Genes and regulation]
Evidence status	1K
Synonyms	**Archive dbSNP** rs59482530, rs17749718
HGVS name	11:g.76301316C>T
Genotyping chips	This variation has assays on **7** chips - click the plus to show

Types of Coding Mutations

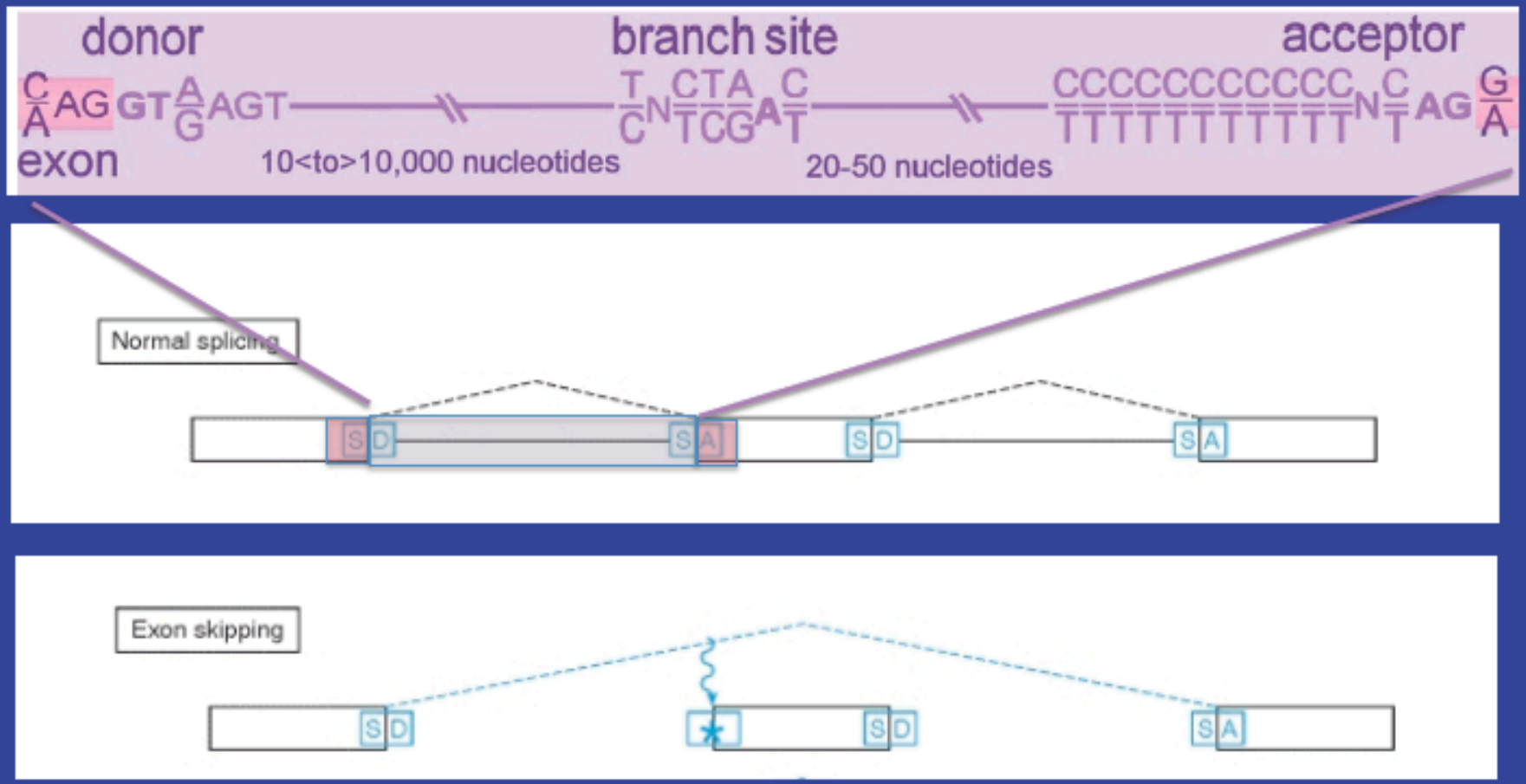
	No mutation	Point mutations			
		Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	**Lys**	**Lys**	**STOP**	**Arg**	**Thr**



basic
polar

K867X **K867R** **K867T**

Splicing Mutations



Courtesy of Laura Elnitski

Example of Nomenclature (NF1)

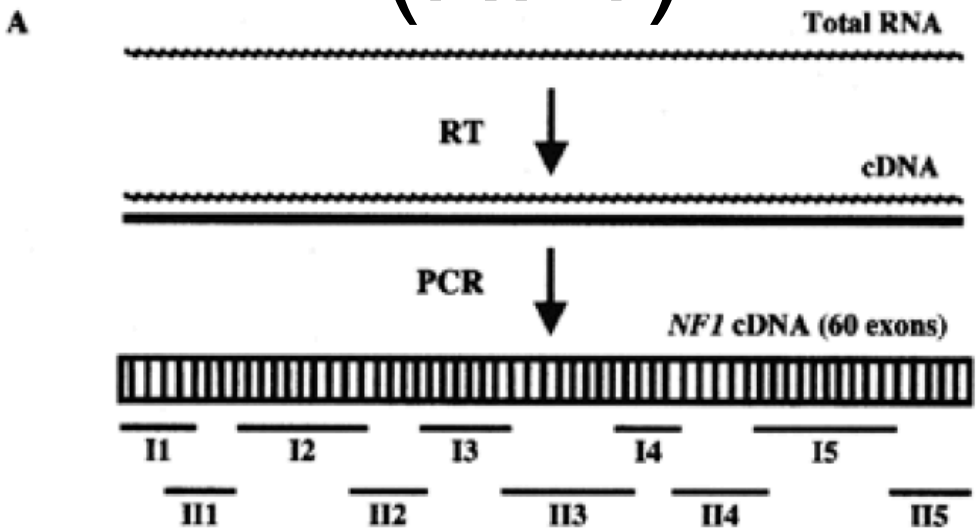


Table 1. *NF1* mutations and their effect on mRNA in 52 unrelated patients with neurofibromatosis type 1

Fragment	Patient	F/S	E/I	Mutation	mRNA level	Type/effect	Protein (aa)	Reference
I1	97-13	F	E 4a	I117S (350T→G)	350U→G	Missense	2818	This report
I1	98-337	S	E 4a	446delA	446delA	Frameshift	163/trunc	This report
I1	96-273	F	I 4b	IVS4b+5G→A	Inact. 5′ ss/skipping E 4b	Splice	163/trunc	This report
I1	97-8	S	E 4b	580delC	580delC	Frameshift	203/trunc	This report
II1	98-397	S	E 4b	527delAT	527delAU	Frameshift	198/trunc	This report
II1	96-16	S	E 5	723insA	723insA	Frameshift	243/trunc	This report
II1	98-1370	S	E 5	717insTCCCACAG	717insUCCCACAG	Frameshift	282/trunc	This report
II1	95-89	F	E 7	910C→T (R304X)	Skipping E 7	Nonsense/splice	2760/IF(–58)	31
I2	94-177	S	E 8	IVS8+1G→A	Inact. 5′ ss/skipping E 8	Splice	2777/IF(–41)	NNFF Consortium
I2	95-137	F	I 10a	IVS10a-9T→A	Inact. 3′ss/cryptic 3′ ss/1392insUUUUUAG	Splice	470/trunc	This report
I2	93-20	S	I 10b	IVS10b+1G→A	Inact. 5′ ss/skipping E 10b	Splice	2773/IF(–45)	55[a]
I2	94-26	S	E 10b	1466A→G (Y489C)	Cryptic 5′ ss/1466del62	Missense/splice	488/trunc	29

Ars *et al.*, HMG 2000

Resource for Sequence Variant Calling

- http://www.hgvs.org/mutnomen/recs.html

HGVS
HUMAN GENOME VARIATION SOCIETY

Recommendations for the description of sequence variants

Last modified March 22, 2016

Since references to WWW-sites are not yet acknowledged as citations, please mention den Dunnen JT and Antonarakis SE (2000). Hum.Mutat. 15:7-12 when referring to these pages.

Contents

Introduction

Discussions regarding the uniform and unequivocal description of sequence variants in DNA and protein sequences (mutations, polymorphisms) were initiated by two papers published in 1993; Beaudet AL & Tsui LC (DOI paper / abstract) and Beutler E (paper / abstract). The original suggestions presented were widely discussed, modified, extended and ultimately resulted in nomenclature recommendations that have been largely accepted and are applied world-wide (*see History*).

evs.gs.washington.edu

Exome Variant Server

NHLBI Exome Sequencing Project (ESP)

Exome Variant Server

Home | Data Browser | Data Usage and Release | How to Use | *What's New* | Contact and FAQ | Downloads

Target: rs71531321 search

examples of valid input for targets (one target per query):

Gene HUGO: ACTB

Gene ID: 60

Chr. Region: 1:1000000-1100000

Single Chr. Location: 7:5567417

rsID: rs71531321

Privacy Terms

evs.gs.washington.edu

Exome Variant Server - Search Results

NHLBI Exome Sequencing Project (ESP)

Exome Variant Server

Variant Results | **Coverage Results**

rsID: rs71531321

Chromosome 7: 5567399 - 5567400

Genes in this region: ACTB(-)

Select Data Set(s)

Check at least one data set below.

Select	Number Variations	Population
☑	1	EuropeanAmerican
☑	0	AfricanAmerican

Display Results

display snp summary →

evs.gs.washington.edu

Exome Variant Server -Variant Table

rsID: rs71531321

Chromosome 7: 5567399 - 5567400

Genes in this region: ACTB(-)

Population: EuropeanAmerican, AfricanAmerican

GWAS Catalog: ACTB

KEGG Pathway: ACTB

Sanger COSMIC: ACTB

PPI STRING 9.0: ACTB

OMIM: ACTB

Variation Color Code:

- splice or nonsense or frameshift
- missense
- coding-synonymous
- coding
- utr
- codingComplex

Download Option:

File Format Text

Zip Format gzip

download

Add or Remove Columns (Description of Columns)

☑ dbSNP rs ID	☑ Alleles	☑ EA Allele Count	☑ AA Allele Count	☑ Allele Count	☑ EA Genotype Count	☑ AA Genotype Count
☑ Genotype Count	☑ MAF (%)	☑ Sample Read Depth	☑ Genes	☑ Gene Accession #	☑ GVS Function	☑ cDNA Change
☑ cDNA Size	☑ Protein Change	☑ Conservation (GERP)	☐ Conservation (phastCons)	☑ Grantham Score	☑ PolyPhen Prediction	☐ Clinical Link
☐ NCBI 37 Allele	☐ Chimp Allele	☐ Illumina HumanExome Chip	☐ GWAS Hits	☐ EA Est. Age (kyrs)	☐ AA Est. Age (kyrs)	☐ GRCh38 Position

Sort Variants by Variant Pos **Select Population** All **Select Transcript** Union of Transcripts

If "Select Transcript" above is set to "Union of Transcripts", and if multiple transcripts of a gene are involved in a variant and the function annotations for the variant are the same, only one representative transcript is listed in the table below for the reasons of speed and space. But annotations for each individual transcript are fully listed in the downloaded file if one chooses to download the data.

Show 10 entries Search:

Variant GRCh37 Pos	rs ID	Alleles	EA Allele #	AA Allele #	All Allele #	EA Genotype #	AA Genotype #	All Genotype #	MAF (%) (EA/AA/All)	Avg. Sample Read Depth	Genes	mRNA Accession #	GVS Function	cDNA Change	cDNA Size	Protein Change	Conservation (GERP)	Grantham Score	PolyPhen2 (Class:Score)
7:5567400	rs71531321	G>A	A=3/G=8597	A=0/G=4406	A=3/G=13003	AA=0/AG=3/GG=4297	AA=0/AG=0/GG=2203	AA=0/AG=3/GG=6500	0.0349/0.0/0.0231	128	ACTB	NM_001101.3	coding-synonymous	c.1107C>T	1128	p.(I369=)	0.86	NA	unknown

Showing 1 to 1 of 1 entries First Previous 1 Next Last

*In general, the INDEL calls are less robust than the SNP calls and have a higher false positive rate. When applying the ESP data to research studies, users are advised to keep this difference in mind.

~ sign in front of a dbSNP rsID indicates an INDEL is approximately mapped to the rsID by SeattleSeqAnnotation137. It should be considered as a suggestion rather than an accurate mapping to the existing records in dbSNP.

http://exac.broadinstitute.org

exac.broadinstitute.org

ExAC Browser

ExAC Browser Beta About Downloads Terms Contact FAQ

ExAC Browser (Beta) | Exome Aggregation Consortium

Search for a gene or variant or region

Examples - Gene: PCSK9, Transcript: ENST00000407236, Variant: 22-46615880-T-C, Multi-allelic variant: rs1800234, Region: 22:46615715-46615880

About ExAC

The Exome Aggregation Consortium (ExAC) is a coalition of investigators seeking to aggregate and harmonize exome sequencing data from a wide variety of large-scale sequencing projects, and to make summary data available for the wider scientific community.

The data set provided on this website spans 60706 unrelated individuals sequenced as part of various disease-specific and population genetic studies. The ExAC Principal Investigators and groups that have contributed data to the current release are listed

Recent News

March 14, 2016

- Version 0.3.1 ExAC data and browser (beta) is released! (Release notes)

January 13, 2015

- Version 0.3 ExAC data and browser (beta) is released! (Release notes)

Genetic Variation: Copy Number

Insertion/deletions of small number of bp (indels):

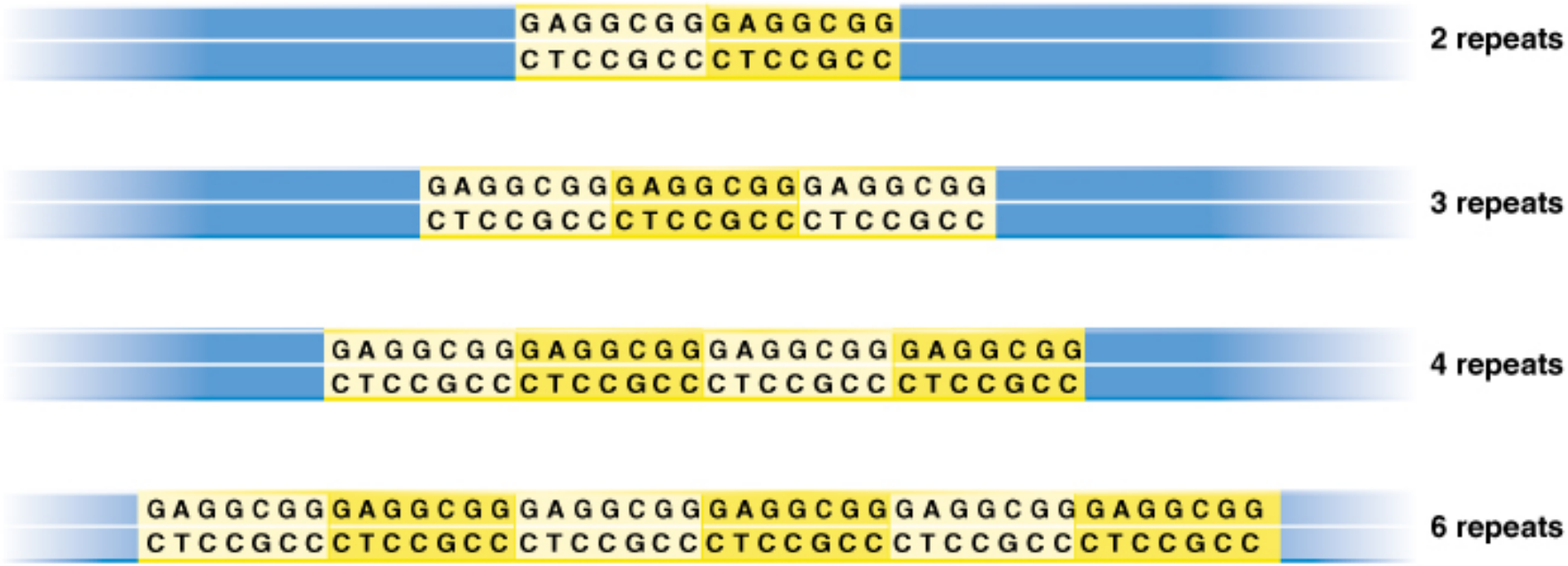
- Polymorphism - also known as copy number variation (CNV)
- Indels in coding region – frameshift mutation if less than 3 bp
- Repeat sequences: di (TG), tri, or tetra – microsatellites

Larger repeats: VNTRs (variable number of tandem repeats), minisatellites - used for forensics

Deletions or duplications of larger blocks of DNA encompassing one or more genes (Williams or DiGeorge Syndrome)

Genetic Variation: Copy Number

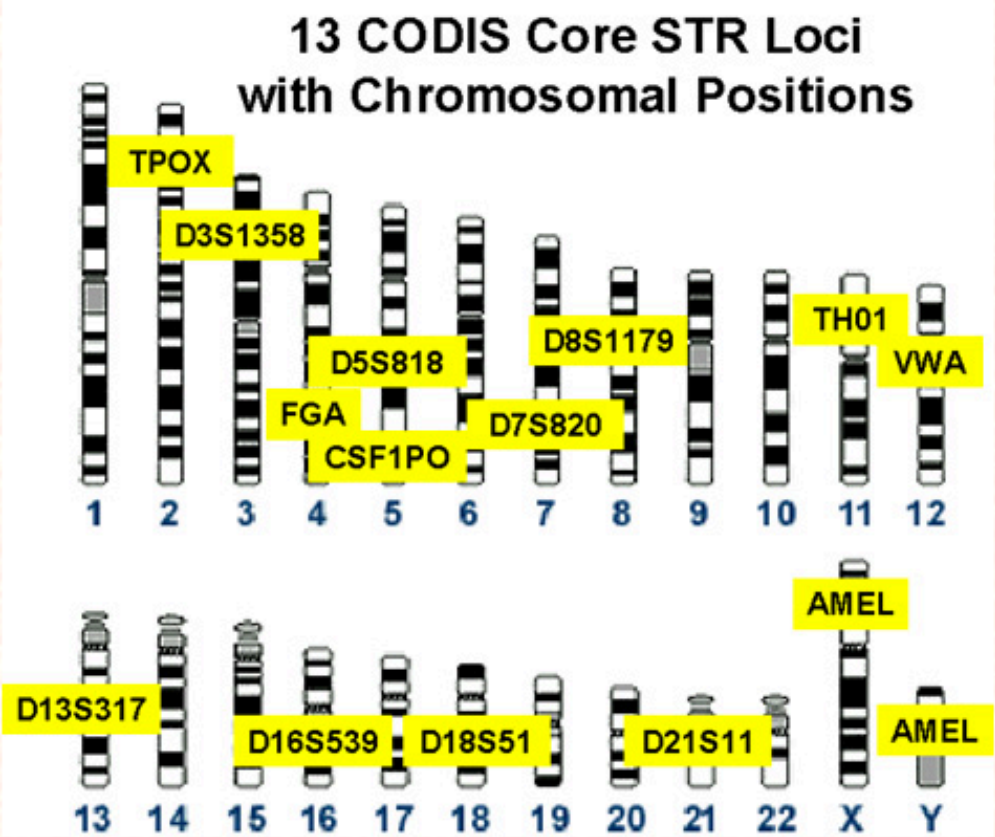
Larger repeats: VNTRs (variable number of tandem repeats), minisatellites - used for forensics



FBI CODIS Core STR Loci:

The FBI has published its thirteen core loci for the Combined DNA Index System (CODIS) database. STR Fact Sheets are available for all thirteen loci (click on locus name for the STR Fact Sheet).
For more information, see: Butler, J.M. (2006) Genetics and genomics of core STR loci used in human identity testing. *J. Forensic Sci.* 51(2): 253-265.

Click on locus name for the appropriate STR Fact Sheet.



Genotype Profiles for six Population Groups at the 13 CODIS Short Tandem Repeat Core Loci and Other PCR-Based Loci (Click on "Link to raw data" at the bottom of the page).

U.S. population data collected by NIST - allele frequencies for 13 CODIS STR loci + D2S1338 and D19S433

Description of diseases and mutations

- Online Mendelian Inheritance in Man (McKusick): http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

omim.org

OMIM Entry Statistics

Home About Statistics Downloads Help External Links Terms of Use Contact Us MIMmatch Donate

Search OMIM... Search

Advanced Search

OMIM Entry Statistics

Number of Entries in OMIM

(Updated March 28th, 2016) :

Prefix	Autosomal	X Linked	Y Linked	Mitochondrial	Totals
* Gene description	14,439	704	48	35	15,226
+ Gene and phenotype, combined	81	2	0	2	85
# Phenotype description, molecular basis known	4,353	303	4	29	4,689
% Phenotype description or locus, molecular basis unknown	1,500	126	5	0	1,631
Other, mainly phenotypes with suspected mendelian basis	1,693	112	2	0	1,807
Totals	22,066	1,247	59	66	23,438

Human mutation database : http://www.hgmd.org

OMIM
Online Mendelian Inheritance in Man
Johns Hopkins University

Characteristics of Simple/ Mendelian diseases

- >20,000 Mendelian traits described in OMIM (Online Mendelian Inheritance in Man)
- Freq usually < 1/10,000
- Most mutations: Loss of function and recessive phenotypes, e.g. inborn errors of metabolism
- Dominant mutations: Cause disease due to 50% of protein product (haploinsufficiency) or dominant negative effect
- Some dominant mutations lead to "gain of function" - e.g. expanded polyglutamine repeat leading to abnormal aggregate (triplet repeat diseases)

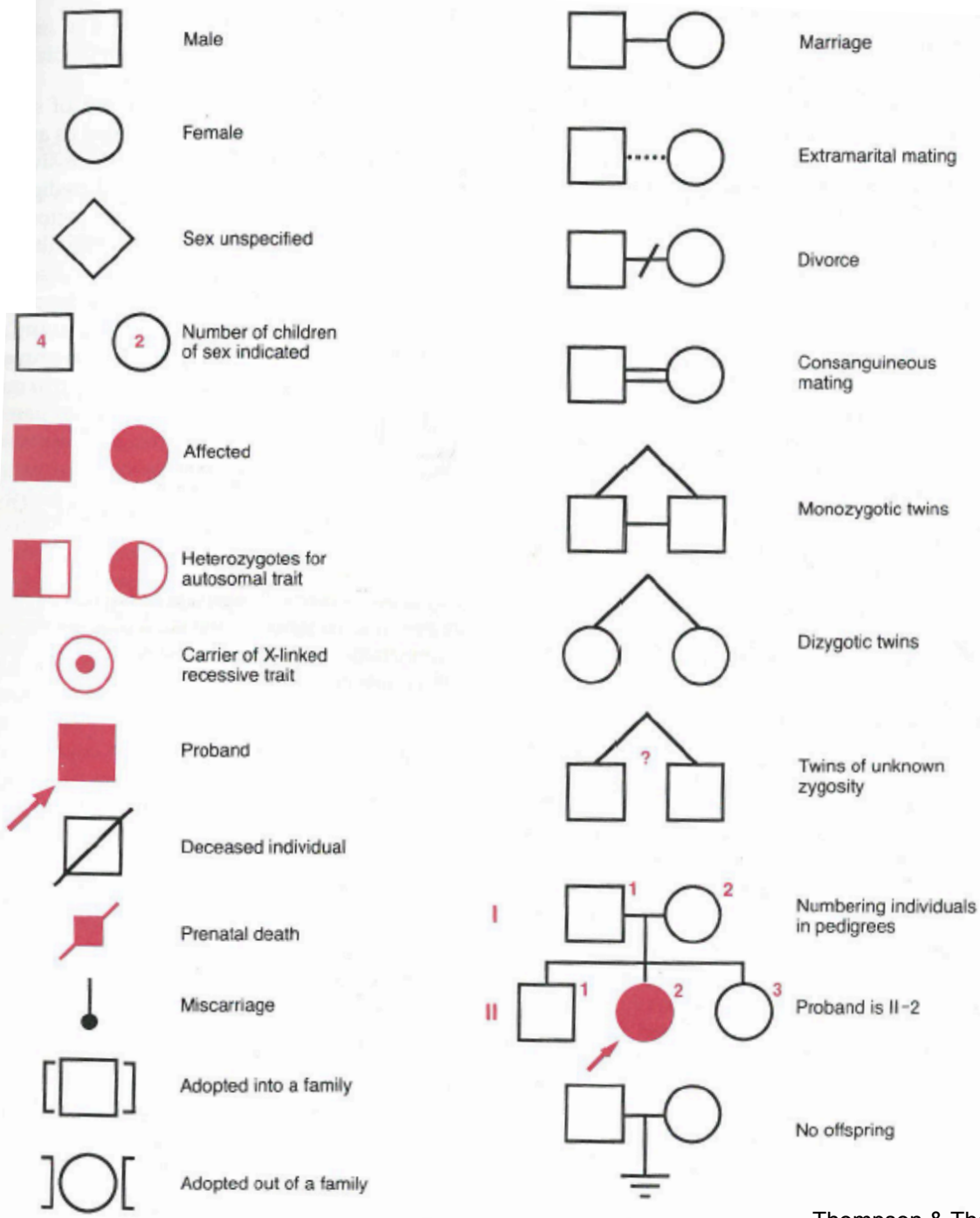


Figure 4-1. Symbols commonly used in pedigree charts.

Thompson & Thompson, Genetics in Medicine (5th)

Basic Mendelian pedigree patterns

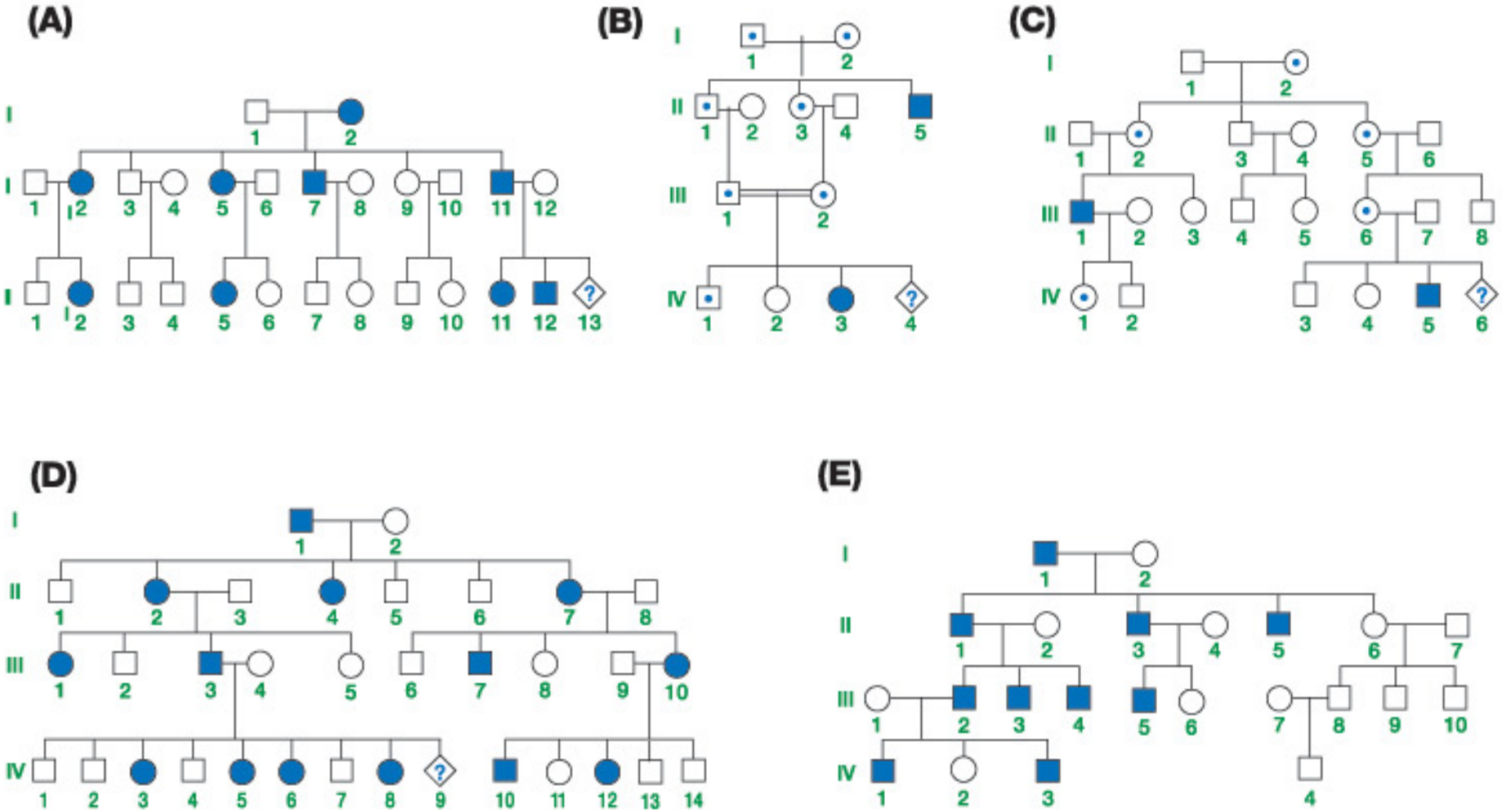
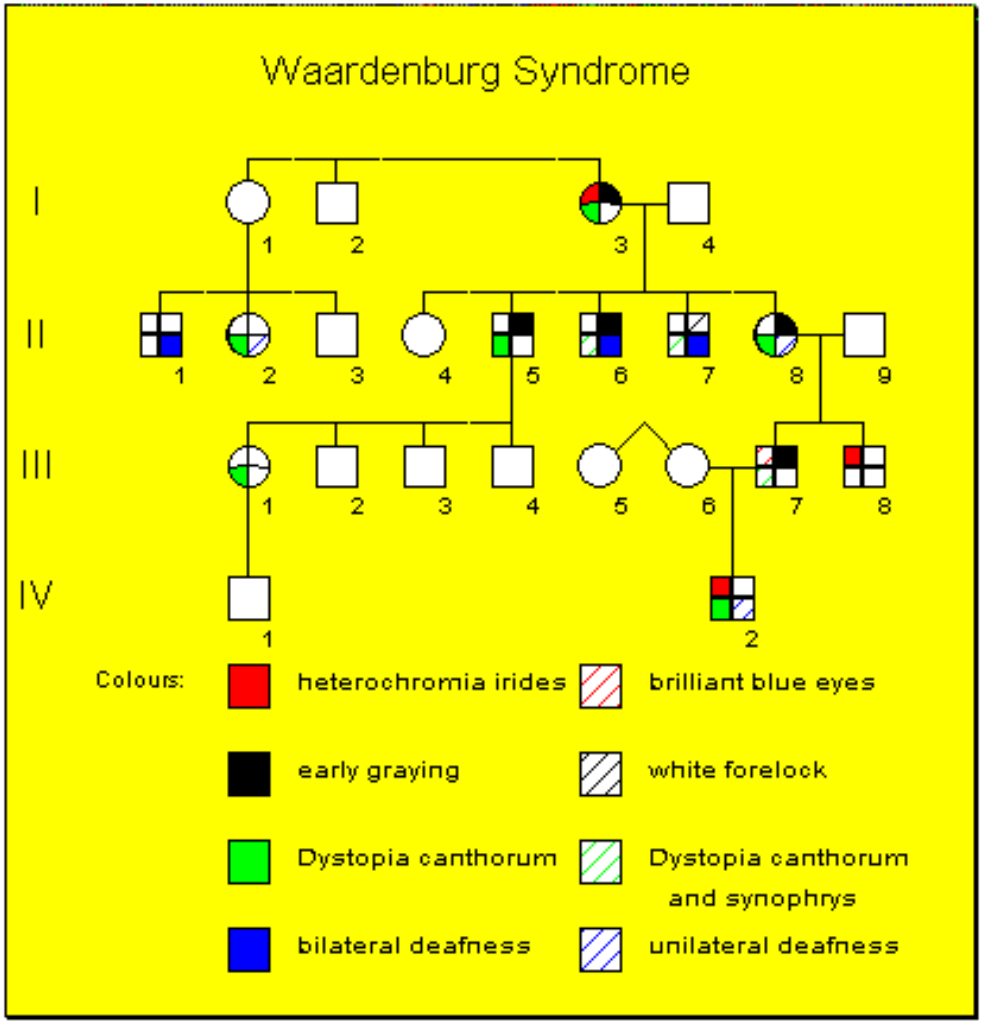
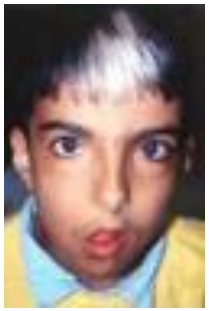


Figure 4-2 Human Molecular Genetics, 3/e. (© Garland Science 2004)

Example of Mendelian disorder: Waardenburg Syndrome (PAX3: Deafness in association with pigmentary anomalies and defects of neural crest-derived tissues).





Loss of function mutations in PAX3 gene: Type 1 Waardenburg syndrome

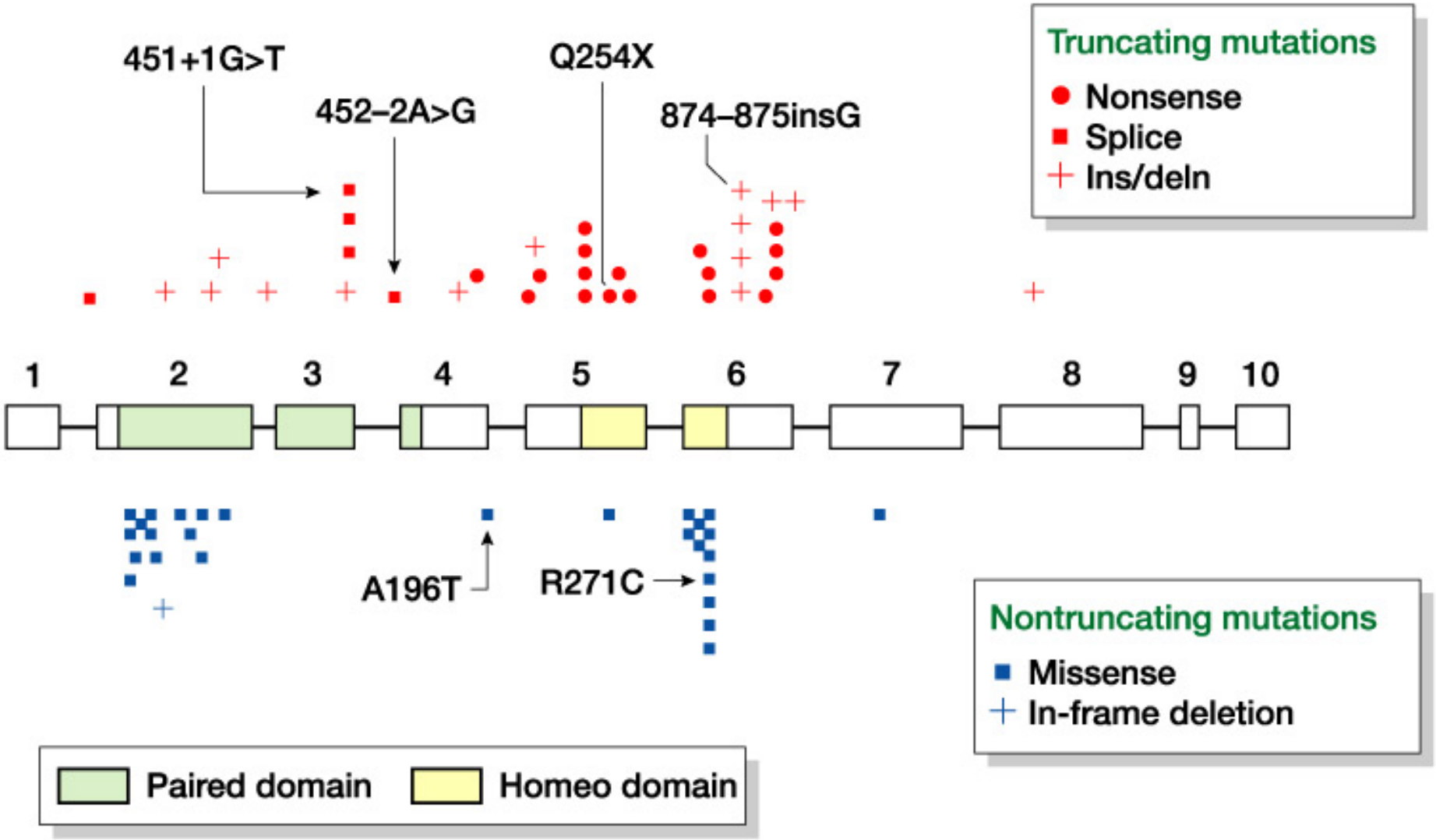
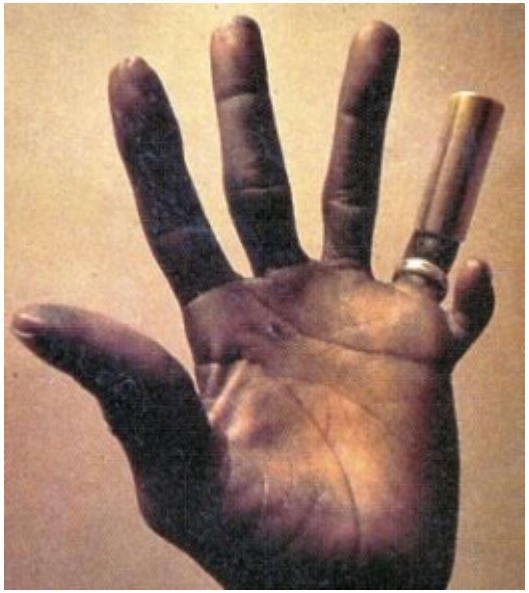


Figure 16-1 Human Molecular Genetics, 3/e. (© Garland Science 2004)

Variable penetrance versus variable expressivity



<u>Penetrance</u> - the frequency of expression of an allele when it is present in the genotype of the organism

Example: if 9/10 of individuals carrying an allele express the trait, the trait is said to be 90% penetrant

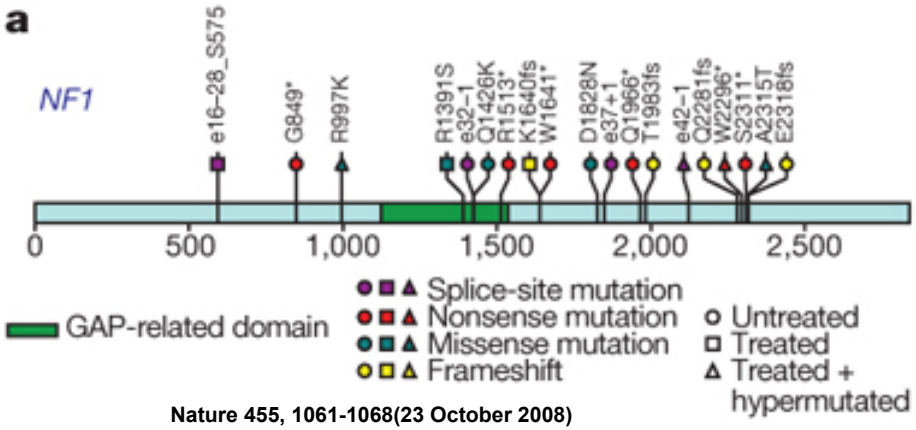
<u>Expressivity</u> - variation in allelic expression when the allele is penetrant.

Example: For polydactyly, an extra digit may occur on one or more appendages, and the digit can be full size or just a stub.

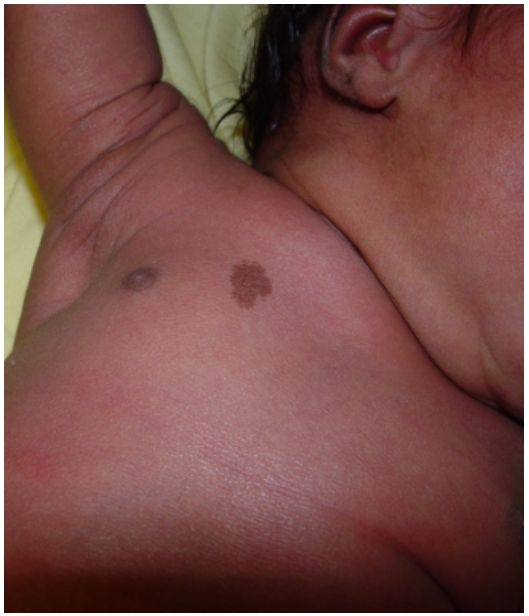
Modifier genes can affect penetrance, dominance, expressivity

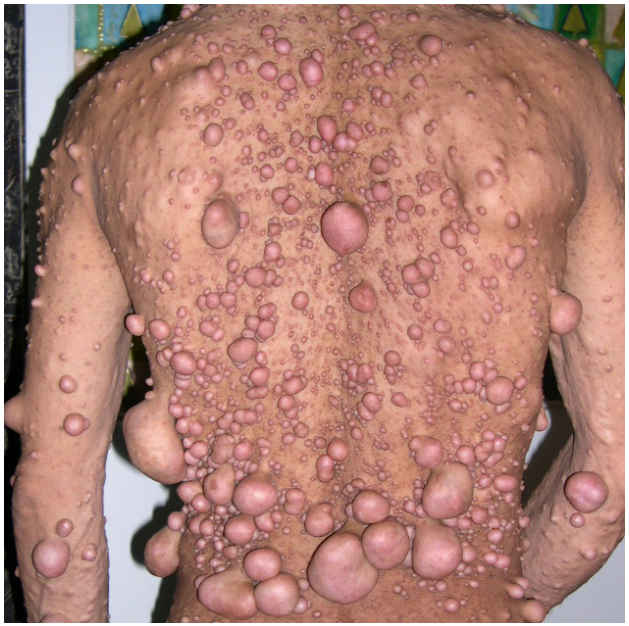
Neurofibromatosis (NF1)

- **Mutations in neurofibromin gene**
- **Reduced penetrance**
- **Wide range of symptoms – variable expressivity**



Nature 455, 1061-1068(23 October 2008)



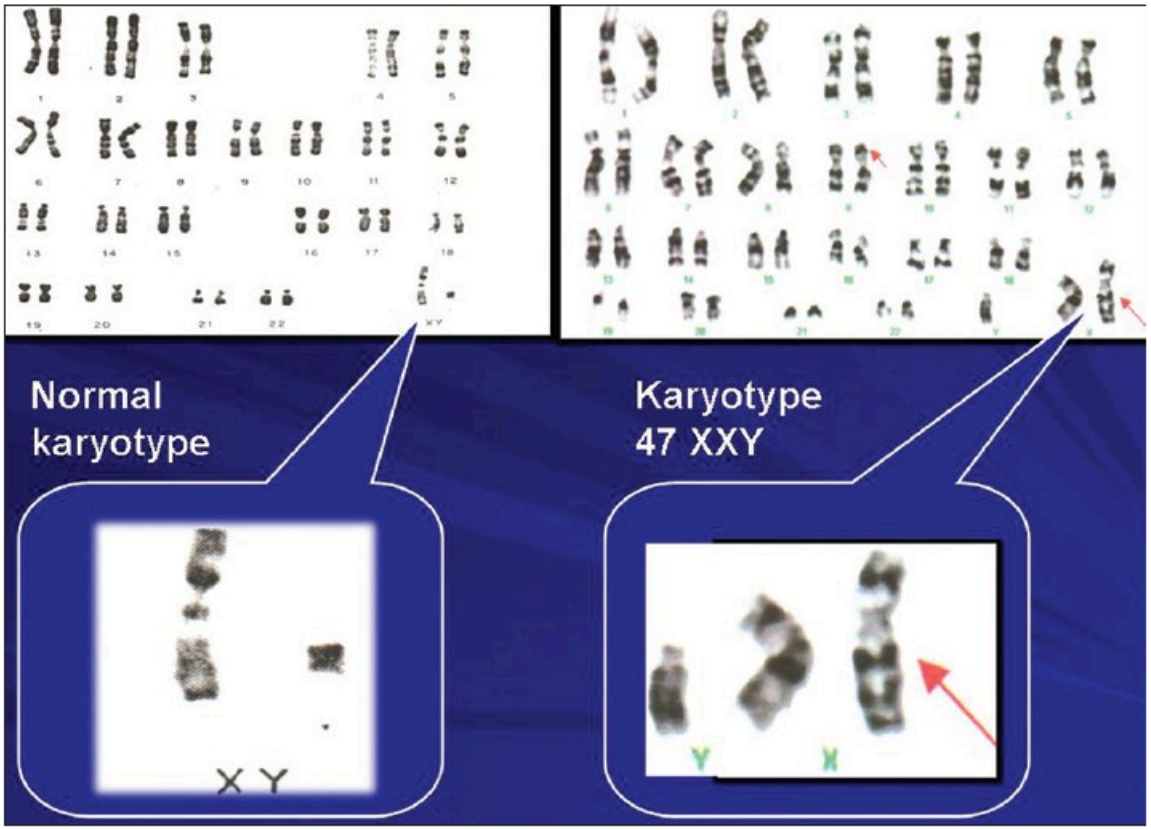


Somatic Mosaicism and Related Syndromes

- Often involve trisomies – nondisjunction event
- Syndrome severity related to percentage of cells
- with aberrant genotype

Klinefelter (47,XXY/46, XY)

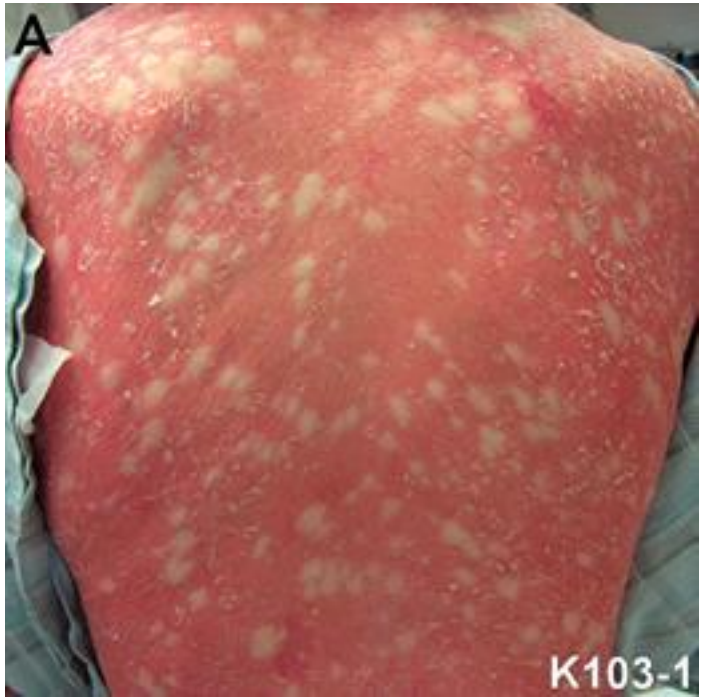
Figure 1: Karyotype XXY



Description of the mosaic form of the Turner syndrome

- 45,X [40] / 46,XX [35]
- 40 + 35 = 75 cells had been examined
- In 40 cells monosomy X was identified.
- 35 cells had normal karyotype.

Ichthyosis with confetti
Revertant mosaicism



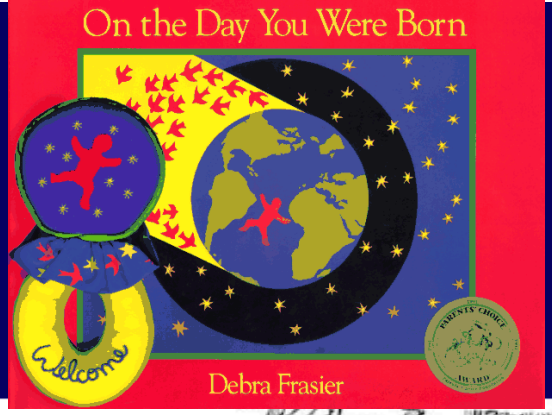
Choate et al., Science 2010

How did we get here?

- ~1,200 “simple” Mendelian trait loci have been cloned, mostly by genetic linkage analysis and positional cloning.
- The revolution in human genetics has been driven by advances in genomics, DNA sequencing and polymorphism detection.

Figure 8-1 Human Molecular Genetics, 3/e. (© Garland Science 2004)

~20 Years Ago

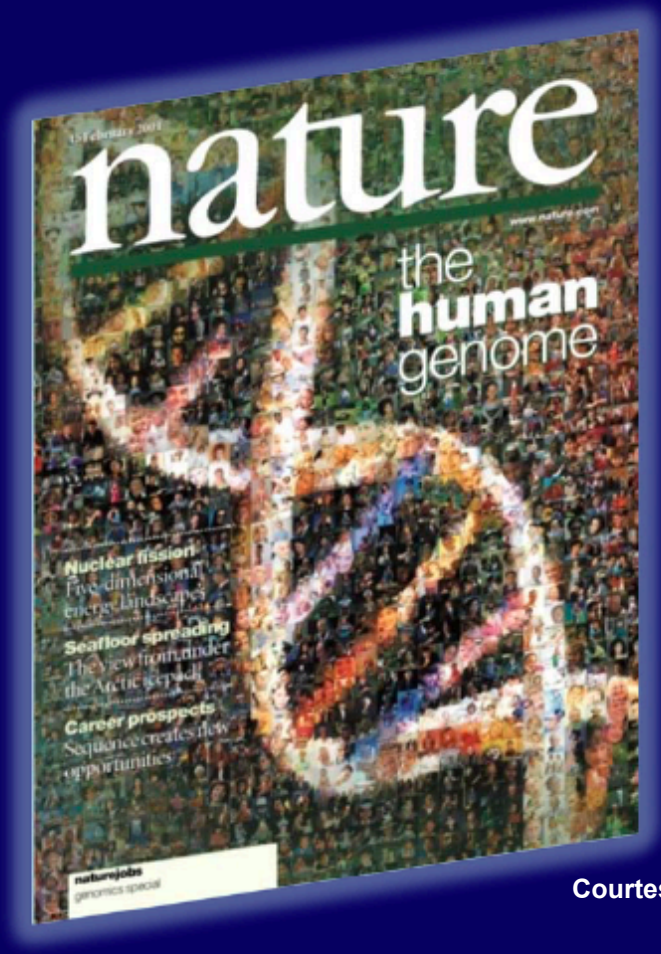


Courtesy of Eric Green

October 1990

Human Genome Project Begins

~10 Years Ago



Courtesy of Eric Green

February 2001

Draft Human Genome Sequence Published

Post HGP

International HapMap Project

International HapMap Project

Home | About the Project | Data | Publications | Tutorial

中文 | English | Français | 日本語 | Yoruba

The International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals. See "**About the International HapMap Project**" for more information.

1000 Genomes

A Deep Catalog of Human Genetic Variation

Home | About | Data | Analysis | Participants | Contact | Browser | Wiki | FTP search | Search

ARTICLE

1000 Genomes Project

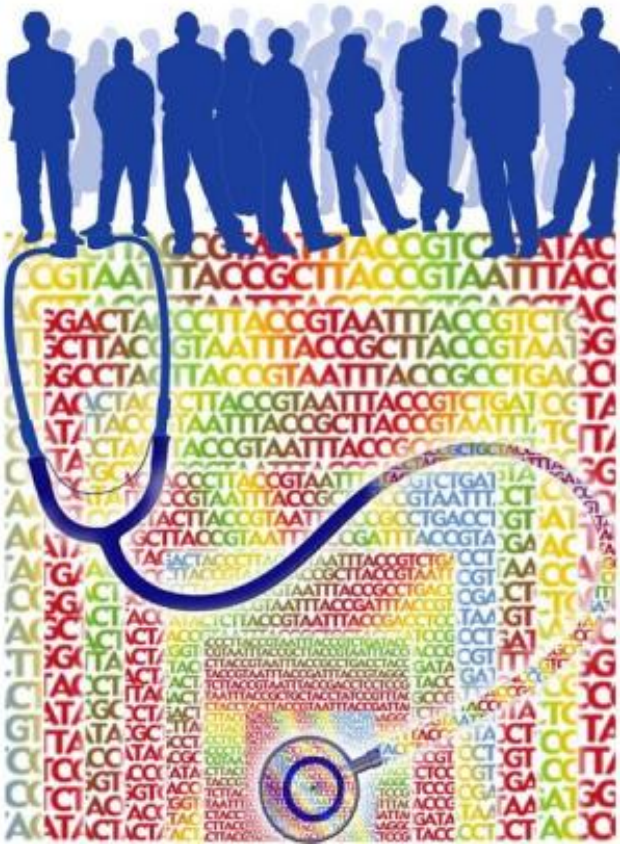
doi:10.1038/nature09534

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

sequence about 2,000 unidentified individuals from 20 populations around the world

Capture minor allele frequencies as low as 1%



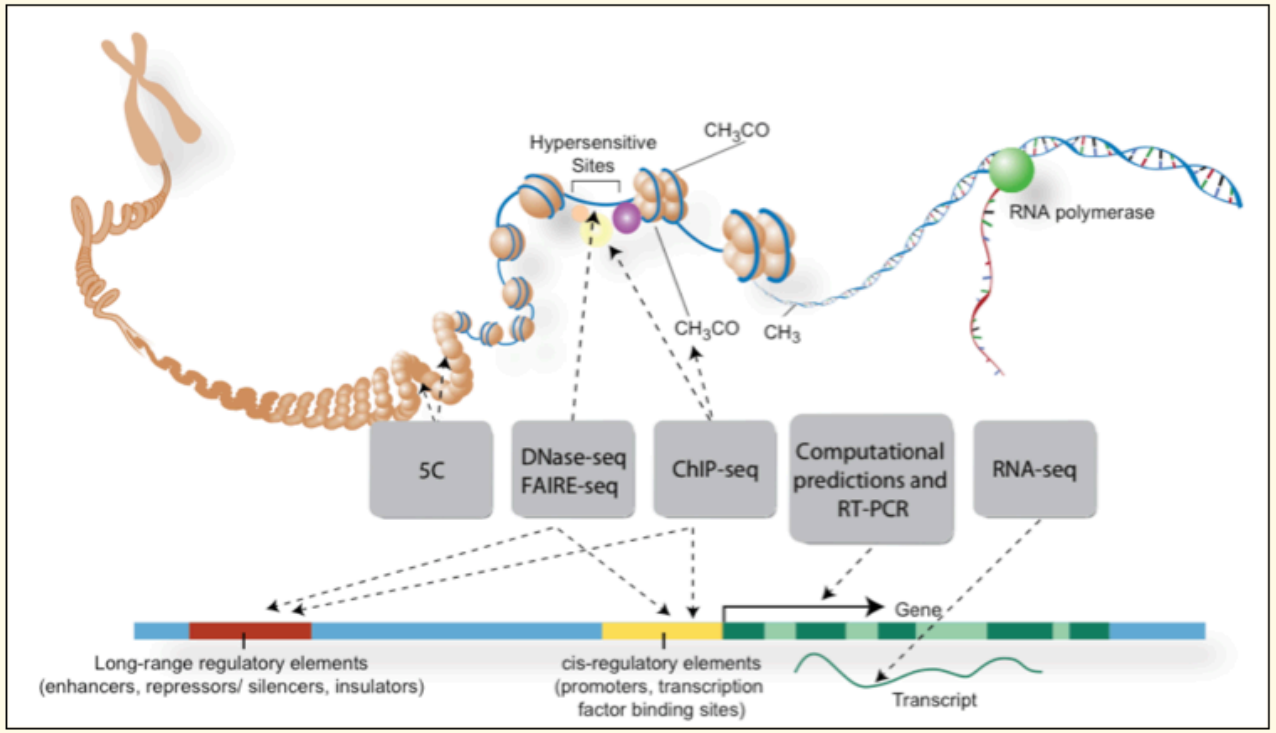
Post HGP

ENCODE Data Coordination Center at UCSC

Home - Downloads - Data Policy - Help

About ENCODE Data at UCSC

ENCODE investigators employ a variety of assays and methods to identify functional elements. The discovery and an elements is accomplished primarily by sequencing RNA from a diverse range of sources, comparative genomics, inte bioinformatic methods, and human curation. Regulatory elements are typically investigated through DNA hypersensiti of DNA methylation, and chromatin immunoprecipitation (ChIP) of proteins that interact with DNA, including modified transcription factors, followed by sequencing (ChIP-Seq).



Credits: Darryl Leja (NHGRI), Ian Dunham (EBI)

Published Genome-Wide Associations
Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories

www.ebi.ac.uk/fgpt/gwas/

GOCI | GWAS Diagram Browser

GWAS Diagram Browser

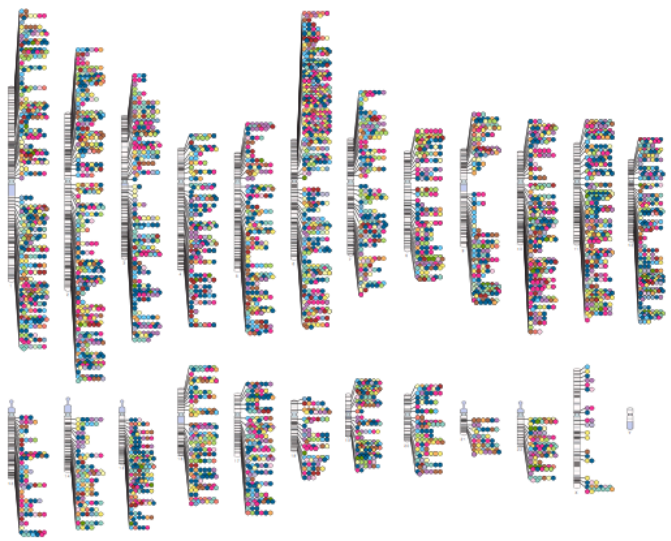
Exploring Genome-wide Association Studies

National Human Genome Research Institute | EMBL-EBI

Query by trait | Clear | *To show only one trait, e.g. "breast cancer" or "schizophrenia", type the trait into the box on the left and hit "Query by trait"*

Interactive GWAS Diagram | Trait-specific Views | Time Series Views | Downloads | Help | Curation process | About | Contact us | Hide Legend

This diagram shows all SNP-trait associations with p-value ≤ 5.0×10^{-8}, published in the GWAS catalogue (http://www.genome.gov/c
information on how to navigate the diagram, see the help tab.



SNP-associated trait categories

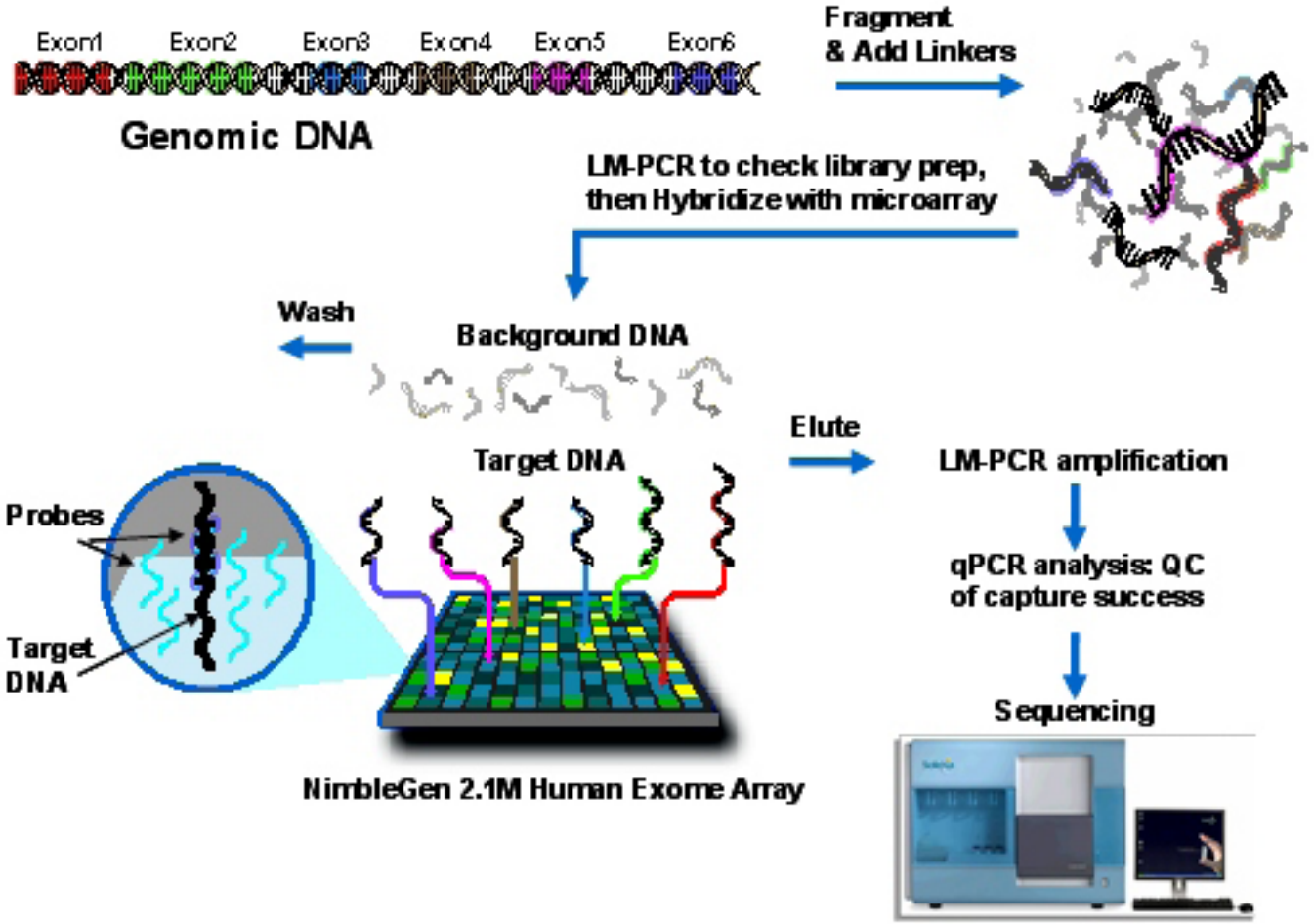
- Digestive system disease
- Cardiovascular disease
- Metabolic disease
- Immune system disease
- Nervous system disease
- Liver enzyme measurement
- Lipid or lipoprotein measurement
- Inflammatory marker measurement
- Hematological measurement
- Body weights and measures
- Cardiovascular measurement
- Other measurement
- Response to drug
- Biological process
- Cancer
- Other disease
- Other trait

- Digestive system disease
- Cardiovascular disease
- Metabolic disease
- Immune system disease
- Nervous system disease
- Liver enzyme measurement
- Lipid or lipoprotein measurement
- Inflammatory marker measurement
- Hematological measurement
- Body measurement
- Cardiovascular measurement
- Other measurement
- Response to drug
- Biological process
- Cancer
- Other disease
- Other trait

NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

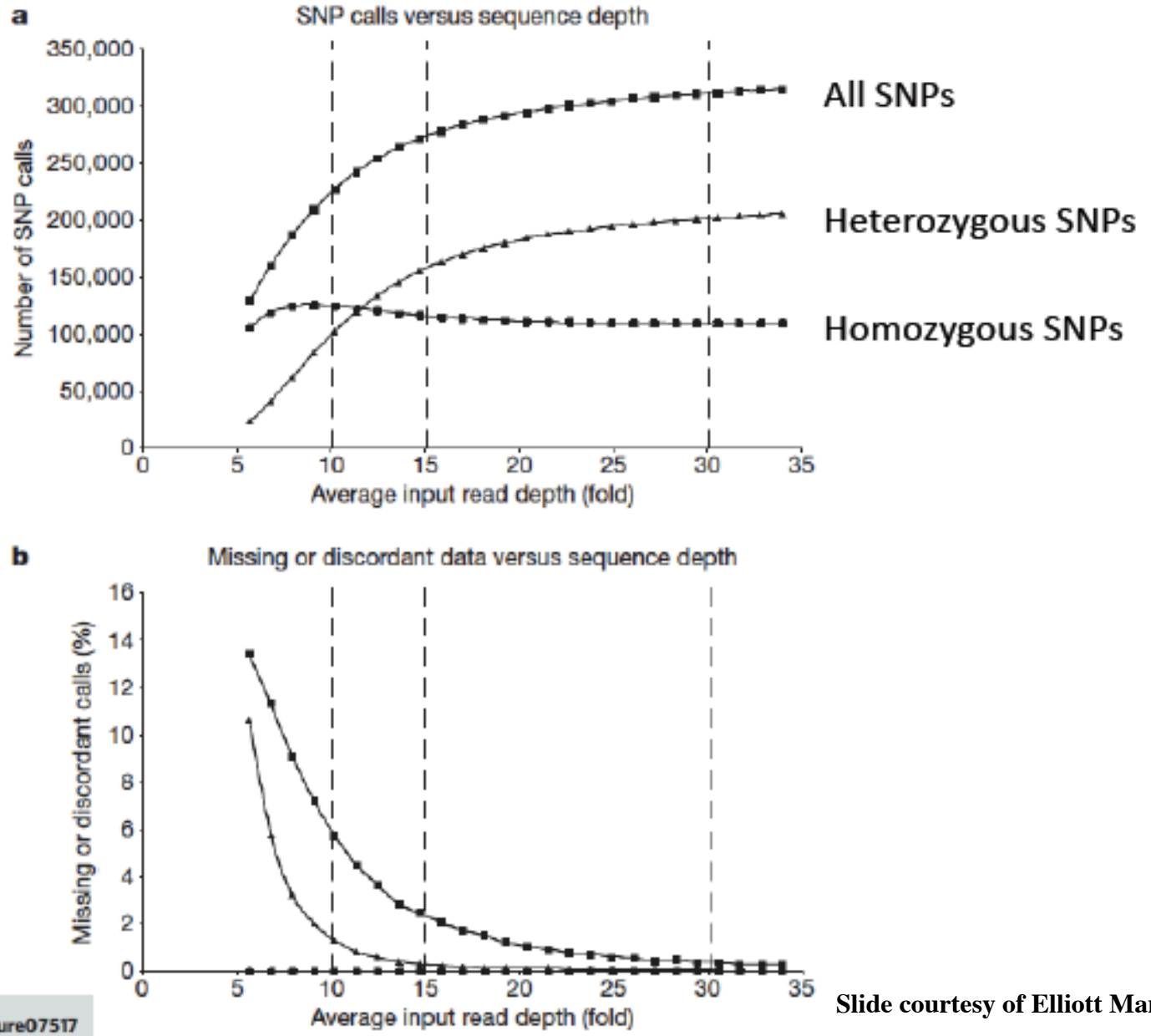
EMBL-EBI



1% of the genome – 30Mb

From Nimblegen

Why 30X?



Vol 456|6 November 2008|doi:10.1038/nature07517

Slide courtesy of Elliott Margulies

doi:10.1038/nature08250

nature

LETTERS

Targeted capture and massively parallel sequencing of 12 human exomes

Sarah B. Ng[1], Emily H. Turner[1], Peggy D. Robertson[1], Steven D. Flygare[1], Abigail W. Bigham[2], Choli Lee[1], Tristan Shaffer[1], Michelle Wong[1], Arindam Bhattacharjee[4], Evan E. Eichler[1,3], Michael Bamshad[2], Deborah A. Nickerson[1] & Jay Shendure[1]

Table 1 | Sequence coverage and array-based validation

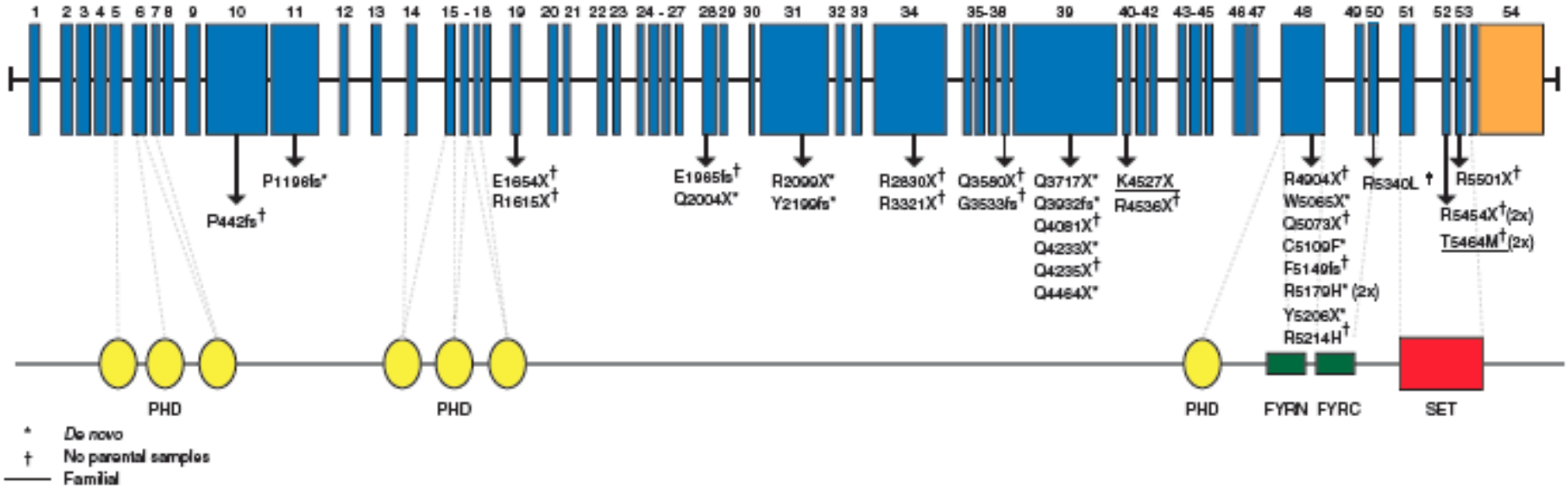
			Concordance with Illumina Human1M-Duo calls		
Individual	Covered ≥1×	Sequence called	Homozygous reference	Heterozygous	Homozygous non-reference
NA18507 (YRI)	26,477,161 (99.7%)	25,795,189 (97.1%)	23757/23762 (99.98%)	5553/5583 (99.46%)	3582/3592 (99.72%)
NA18517 (YRI)	26,476,761 (99.7%)	25,748,289 (97.0%)	23701/23705 (99.98%)	5575/5601 (99.54%)	3568/3579 (99.69%)
NA19129 (YRI)	26,491,035 (99.8%)	25,733,587 (96.9%)	23701/23708 (99.97%)	5482/5510 (99.49%)	3681/3690 (99.76%)
NA19240 (YRI)	26,486,481 (99.7%)	25,576,517 (96.3%)	23546/23551 (99.98%)	5600/5634 (99.40%)	3542/3549 (99.80%)
NA18555 (CHB)	26,475,665 (99.7%)	25,529,861 (96.1%)	23980/23984 (99.98%)	4877/4893 (99.67%)	3776/3786 (99.74%)
NA18956 (JPT)	26,454,942 (99.6%)	25,683,248 (96.7%)	24217/24221 (99.98%)	4890/4910 (99.59%)	3751/3760 (99.76%)
NA12156 (CEU)	26,476,155 (99.7%)	25,360,704 (95.5%)	23789/23794 (99.98%)	5493/5514 (99.62%)	3206/3213 (99.78%)
NA12878 (CEU)	26,439,953 (99.6%)	25,399,572 (95.6%)	23885/23891 (99.97%)	5413/5425 (99.78%)	3274/3292 (99.45%)
FSS10066 (Eur)	26,467,140 (99.7%)	25,546,738 (96.2%)	NA	NA	NA
FSS10208 (Eur)	26,461,768 (99.6%)	25,576,256 (96.3%)	NA	NA	NA
FSS22194 (Eur)	26,426,401 (99.5%)	25,454,551 (95.9%)	NA	NA	NA
FSS24895 (Eur)	26,478,775 (99.7%)	25,602,677 (96.4%)	NA	NA	NA

The number of coding bases covered at least 1× and with sufficient coverage to variant call (≥8× and consensus quality ≥30) are listed for each exome, with the fraction of the aggregate target (26.6 Mb) that this represents in parentheses. For the eight HapMap individuals, concordance with array genotyping (Illumina Human1M-Duo) is listed for positions that are homozygous for the reference allele, heterozygous or homozygous for the non-reference allele (according to the array genotype). CEU, CEPH HapMap; CHB, Chinese HapMap; Eur, European–American ancestry (non-HapMap); JPT, Japanese HapMap; YRI, Yoruba HapMap. NA, Not applicable.

nature genetics

Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome

Sarah B Ng[1,7], Abigail W Bigham[2,7], Kati J Buckingham[2], Mark C Hannibal[2,3], Margaret J McMillin[2], Heidi I Gildersleeve[2], Anita E Beck[2,3], Holly K Tabor[2,3], Gregory M Cooper[1], Heather C Mefford[2], Choli Lee[1], Emily H Turner[1], Joshua D Smith[1], Mark J Rieder[1], Koh-ichiro Yoshiura[4], Naomichi Matsumoto[5], Tohru Ohta[6], Norio Niikawa[6], Deborah A Nickerson[1], Michael J Bamshad[1–3] & Jay Shendure[1]

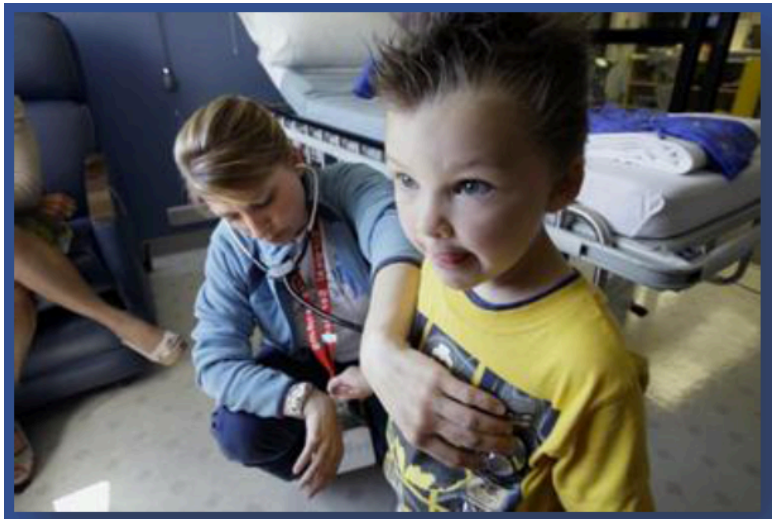


Genetics in Medicine • Volume 13, Number 3, March 2011

BRIEF REPORT

Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease

Elizabeth A. Worthey, PhD[1,2], Alan N. Mayer, MD, PhD[2,3], Grant D. Syverson, MD[2], Daniel Helbling, BSc[1], Benedetta B. Bonacci, MSc[2], Brennan Decker, BSc[1], Jaime M. Serpe, BSc[2], Trivikram Dasu, PhD[2], Michael R. Tschannen, BSc[1], Regan L. Veith, MSc[2], Monica J. Basehore, PhD[4], Ulrich Broeckel, MD, PhD[1,2,3], Aoy Tomita-Mitchell, PhD[1,2,3], Marjorie J. Arca, MD[3,5], James T. Casper, MD[2,3], David A. Margolis, MD[2,3], David P. Bick, MD[1,2,3], Martin J. Hessner, PhD[1,2], John M. Routes, MD[2,3], James W. Verbsky, MD, PhD[2,3], Howard J. Jacob, PhD[1,2,3,6], and David P. Dimmock, MD[1,2,3]



Credit: Gary Porter, Milwaukee Journal Sentinel

Exome sequencing – *XIAP*
Whole Bone Marrow Transplant

January 2010

The Human Genome… by the Numbers

~5% of Human Genome Sequence is Constrained Across Mammals (and Presumed Functional)

- 5% of 3B Bases = ~150M Bases
- Do NOT Yet Know the Position of these ~150M Functional Bases
- Lower Bound for the Amount that is Functional

~1.5% Encodes for Protein (Genes)

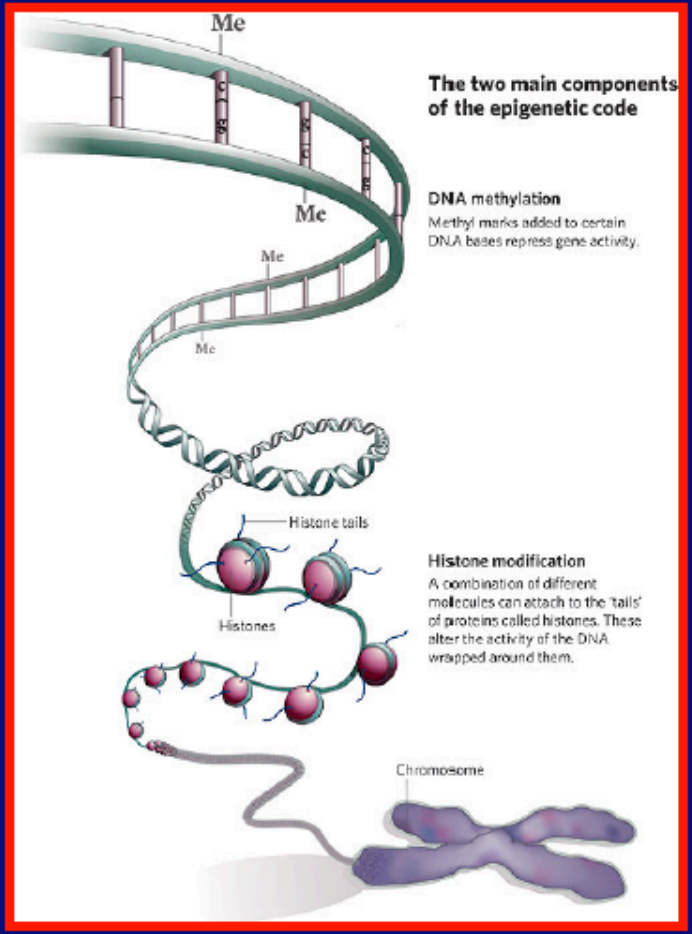
- Corresponds to ~18-22K Genes
- Many More than ~22K Different Proteins
- Good Inventory at Present

~3.5% Functional But Non-Coding

- Gene Regulatory Elements
- Chromosomal Functional Elements
- Undiscovered Functional Elements (NOT Yet in Textbooks!)
- Poor Inventory at Present

Slide courtesy of Eric Green

The Epigenetic Landscape



HEP
Human
Epigenome
Project

ENCODE

Slide courtesy of Eric Green (modified)

ENCODE Project

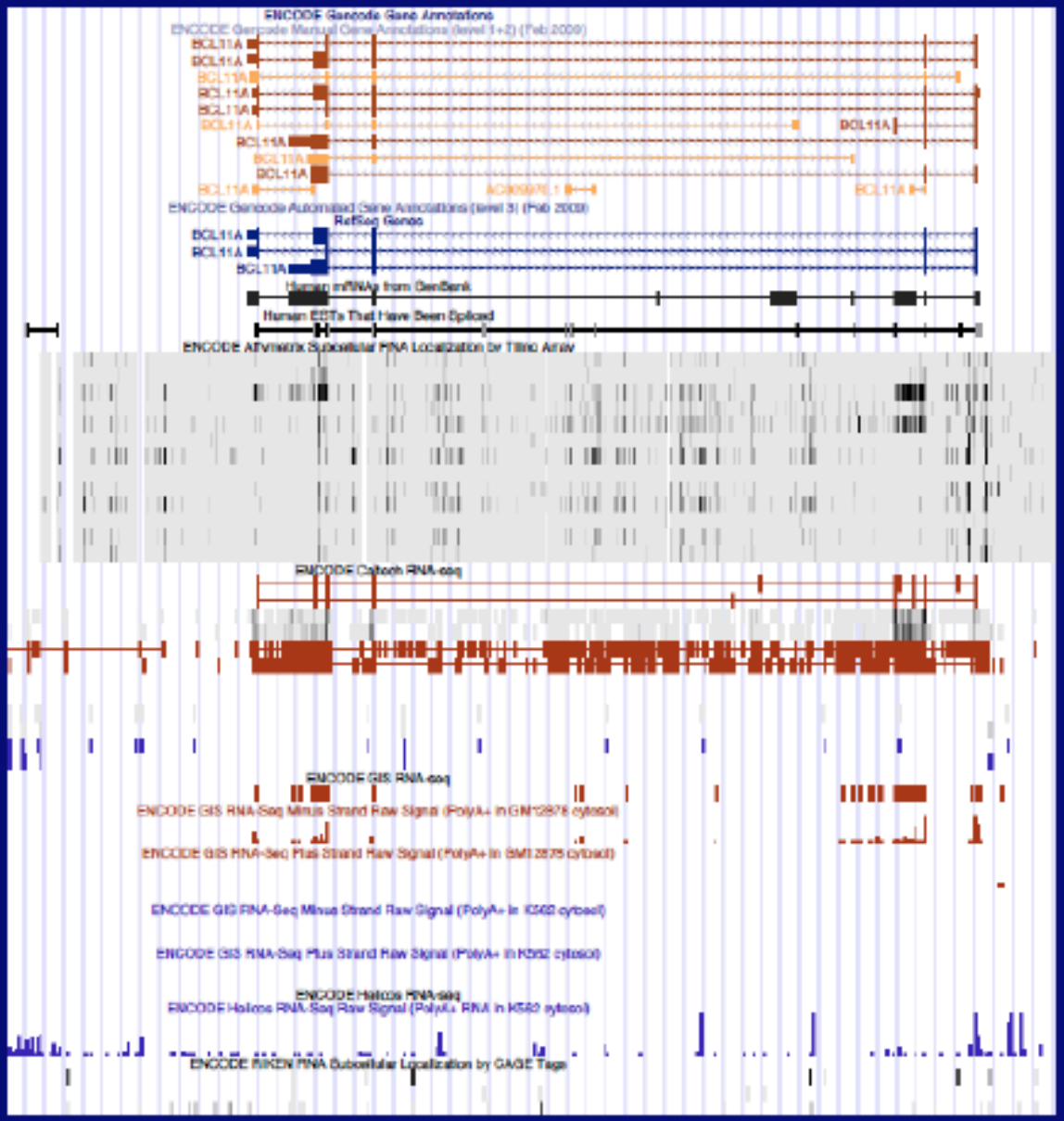
- ENCODE: ENCyclopedia Of DNA Elements
- Goal: Compile a *Comprehensive Encyclopedia* of All Functional Elements in the Human Genome
- Initial Pilot Project: 1% of Human Genome
- Apply Multiple, Diverse Approaches to Study and Analyze that 1% in a Consortium Fashion

Slide courtesy of Eric Green

ENCODE: Lots of Data and Data Types

Generated by:

RNA-Seq
RNA-array
TF ChIP-Seq
Histone modif ChIP-Seq
DNaseHS-Seq
FAIRE-Seq
Methyl-Seq
Methyl27-bisulfite
etc.
etc.
etc.



Slide courtesy of Eric Green

Hon, Hawkins, Ren review (Human Molecular Genetics 2009)

CTCF
- Insulator

H3K4me1, also P300
- Enhancer

H3K4me2
- Repressor? Recruit HDAC

HeK4me3
- Promoter, trithorax (open chromatin)

H3K9ac
H3K9me1
- Activated

H3K27ac
H3K27me3
- Transcriptional elongation?
- Silencing (Plath, Science 2003) – silencing of the X chromosome, PCG to trimethylate H3K27

H3K36me3
- Activated

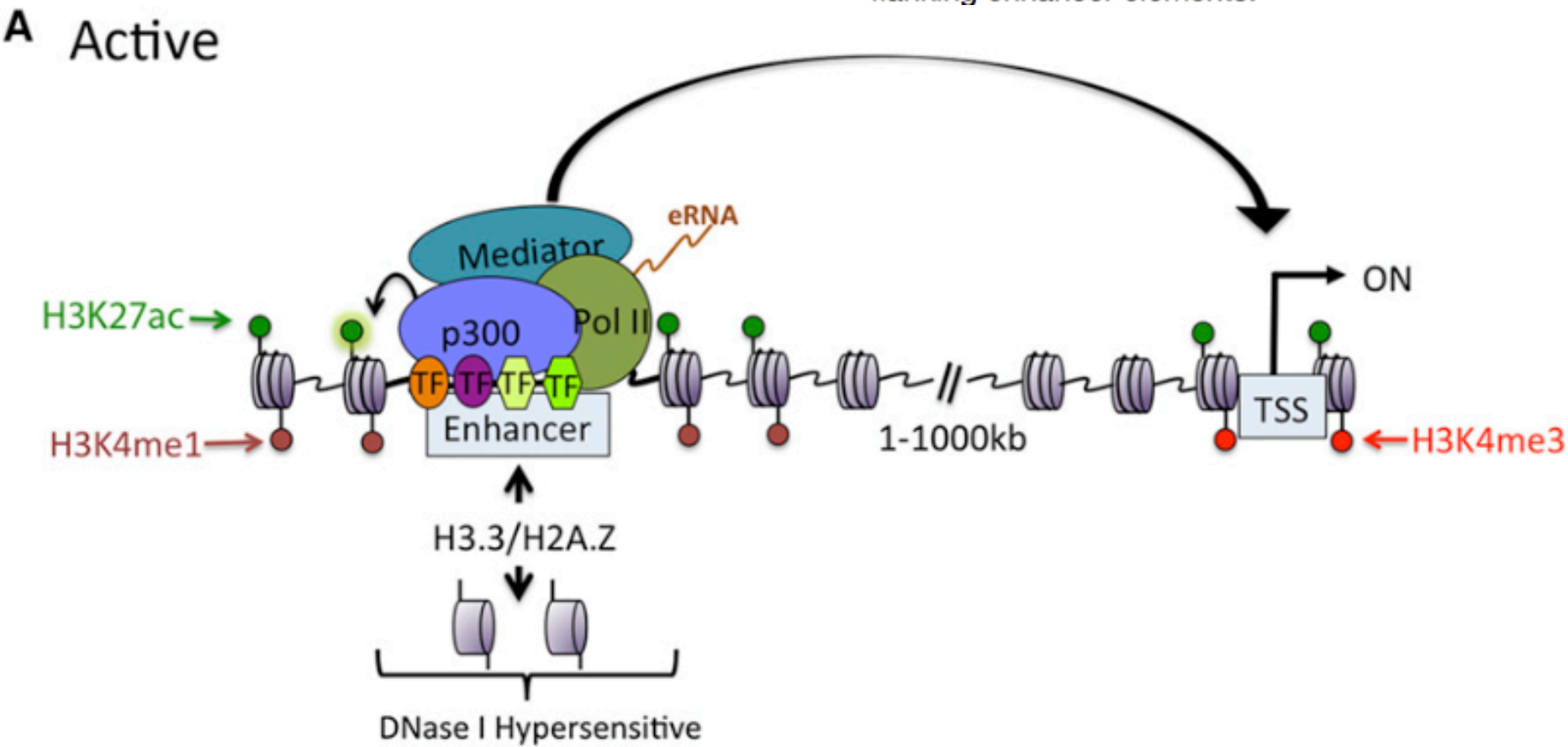
H4K20me1
Pol2(b)

H3.3/H2A.Z mark double variant nucleosome that are unstable and mark nucleosome-free regions in active promoters and regulatory regions

Figure 1. Epigenetic Features of Active, Primed, and Poised Enhancers

(A) Schematic representation of the major chromatin features found at active enhancers. Enhancers are associated with incorporation of hypermobile nucleosomes containing H3.3/H2A.Z histone variants, which compete for DNA binding with TFs. TFs in turn recruit coactivator proteins that can modify and remodel nucleosomes. H3K4me1 and H3K27ac are the predominant histone modifications deposited at nucleosomes flanking enhancer elements.

Calo & Wysocka, Mol Cell 2013



nature | ENCODE

Search ENCODE Go

Home | Research | Threads | Additional Research | News and Comment | About | Sponsor

nature ENCODE explorer

THREADS

01 02 03 04 05 06 07 08 09 10 11 12 13

ENHANCER DISCOVERY AND CHARACTERIZATION

ENCODE has developed methods to discover enhancers, and characterized them using comparisons with other data sets and by molecular biology experiments

PRODUCED WITH SUPPORT FROM illumina

Welcome to the nature ENCODE explorer

Access the collected papers by exploring the thematic threads that run through them, with topics such as DNA methylation, RNA or machine learning.

Select a thread to start

http://www.nature.com/encode/#/threads

What is ENCODE? | Threads: a new approach | Guide to the ENCODE explorer

ENCODE Project Writes Eulogy for Junk DNA

Science 7 September 2012:
vol. 337 no. 6099 1159–1161

ENCODE By the Numbers

147 cell types studied

80% functional portion of human genome

20,687 protein-coding genes

18,400 RNA genes

1640 data sets

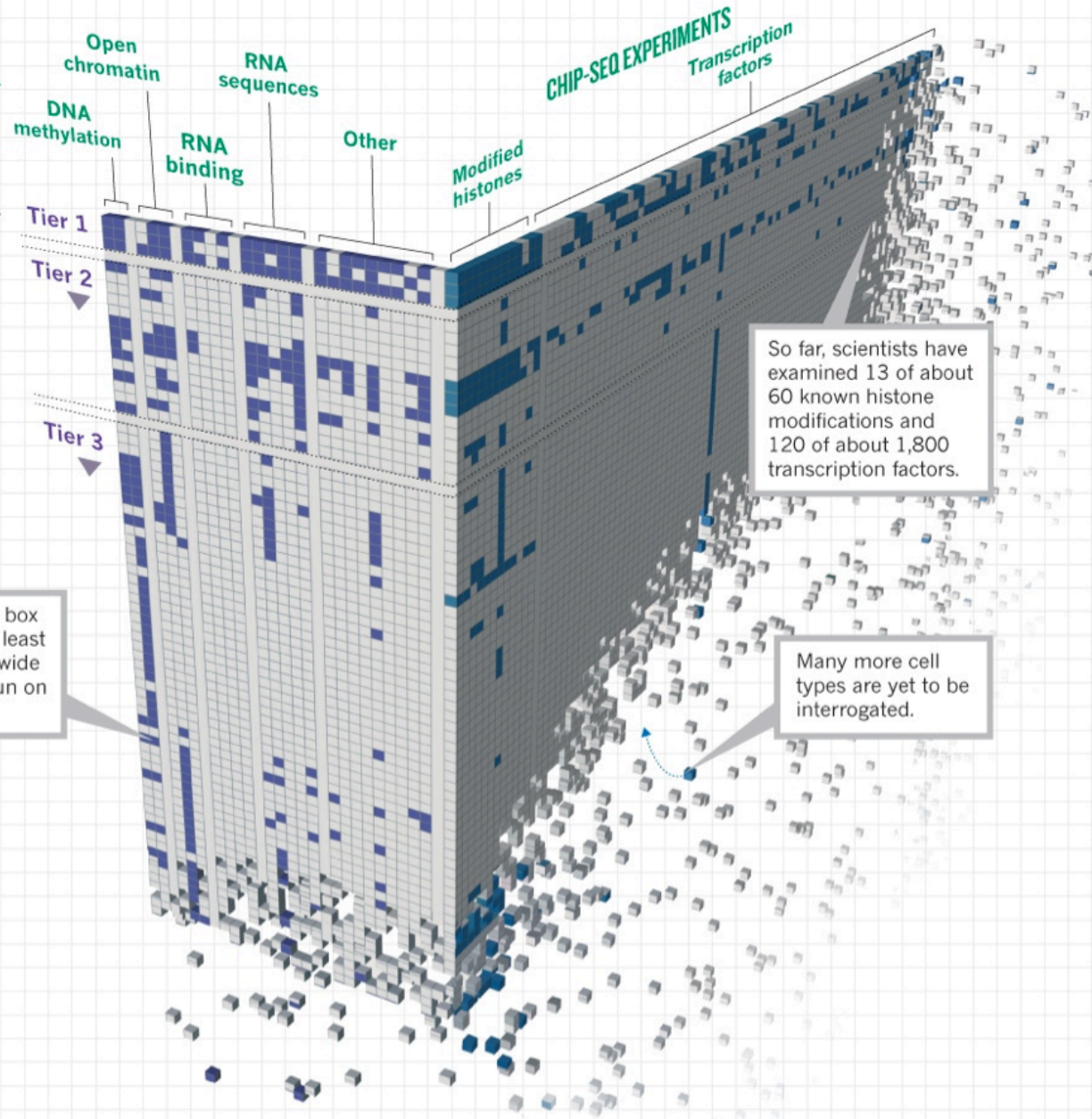
30 papers published this week

442 researchers

$288 million funding for pilot, technology, model organism, and current project

MAKING A GENOME MANUAL

Scientists in the Encyclopedia of DNA Elements Consortium have applied 24 experiment types (across) to more than 150 cell lines (down) to assign functions to as many DNA regions as possible — but the project is still far from complete.



EXPERIMENTAL TARGETS

DNA methylation: regions layered with chemical methyl groups, which regulate gene expression.

Open chromatin: areas in which the DNA and proteins that make up chromatin are accessible to regulatory proteins.

RNA binding: positions where regulatory proteins attach to RNA.

RNA sequences: regions that are transcribed into RNA.

ChIP-seq: technique that reveals where proteins bind to DNA.

Modified histones: histone proteins, which package DNA into chromosomes, modified by chemical marks.

Transcription factors: proteins that bind to DNA and regulate transcription.

CELL LINES

Tiers 1 and 2: widely used cell lines that were given priority.

Tier 3: all other cell types.

Maher, Nature 2012

nature International weekly journal of science

SPECIAL

See all specials

MOUSE ENCODE

The aim of the Encyclopedia of DNA Elements (ENCODE) project was to map functional elements of the genome. The project initially focused on the human genome but was expanded to include the mouse, which has a premier role as a model organism in biomedical research. In this Online Special, *Nature* brings together a selection of articles from across publishing houses that collectively describe the key datasets and analyses of the Mouse ENCODE project. The results of the ENCODE project constitute an invaluable resource to the scientific community. Many of the papers collected below also provide a comparison of these functional genome maps, thereby revealing key similarities as well as differences between the mouse and the human.

Credit: Kelly Krause/Nature

▼ Nature ▼ Science ▼ PNAS ▼ Genome research ▼ Blood

NATURE

A Comparative Encyclopedia of DNA Elements in the Mouse Genome
The mouse ENCODE consortium *et al.*
Nature **515**, 355–364 (19 November 2014)

Journal home | Subscribe
Current issue | E-alert sign up
For authors | RSS feed

Most read

Mutations in regulatory regions

- Altered TF binding to promoter or enhancer

  e.g. sonic hedgehog enhancer
- Splice site alterations (e.g. some beta thalassemias)
- Altered mRNA stability (e.g. in AAUAA)
- Altered micro RNA binding leading to altered protein translation

  Sequence variants in SLITRK1 are associated with Tourette's syndrome. Science 2005, 310.

A common sex-dependent mutation in a *RET* enhancer underlies Hirschsprung disease risk

Eileen Sproat Emison[1]*, Andrew S. McCallion[1]*, Carl S. Kashuk[1], Richard T. Bush[1], Elizabeth Grice[1], Shin Lin[1], Matthew E. Portnoy[2], David J. Cutler[1], Eric D. Green[2,3] & Aravinda Chakravarti[1]

[1]*McKusick – Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA*
[2]*Genome Technology Branch and* [3]*NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA*

*These authors contributed equally to this work

The identification of common variants that contribute to the genesis of human inherited disorders remains a significant challenge. Hirschsprung disease (HSCR) is a multifactorial, non-mendelian disorder in which rare high-penetrance coding sequence mutations in the receptor tyrosine kinase RET contribute to risk in combination with mutations at other genes. We have used family-based association studies to identify a disease interval, and integrated this with comparative and functional genomic analysis to prioritize conserved and functional elements within which mutations can be sought. We now show that a common non-coding *RET* variant within a conserved enhancer-like sequence in intron 1 is significantly associated with HSCR susceptibility and makes a 20-fold greater contribution to risk than rare alleles do. This mutation reduces *in vitro* enhancer activity markedly, has low penetrance, has different genetic effects in males and females, and explains several features of the complex inheritance pattern of HSCR. Thus, common low-penetrance variants, identified by association studies, can underlie both common and rare diseases.

ARTICLE Nature March 23 2011 Advance issue
doi:10.1038/nature09906

Mapping and analysis of chromatin state dynamics in nine human cell types

Jason Ernst[1,2], Pouya Kheradpour[1,2], Tarjei S. Mikkelsen[1], Noam Shoresh[1], Lucas D. Ward[1,2], Charles B. Epstein[1], Xiaolan Zhang[1], Li Wang[1], Robbyn Issner[1], Michael Coyne[1], Manching Ku[1,3,4], Timothy Durham[1], Manolis Kellis[1,2]* & Bradley E. Bernstein[1,3,4]*

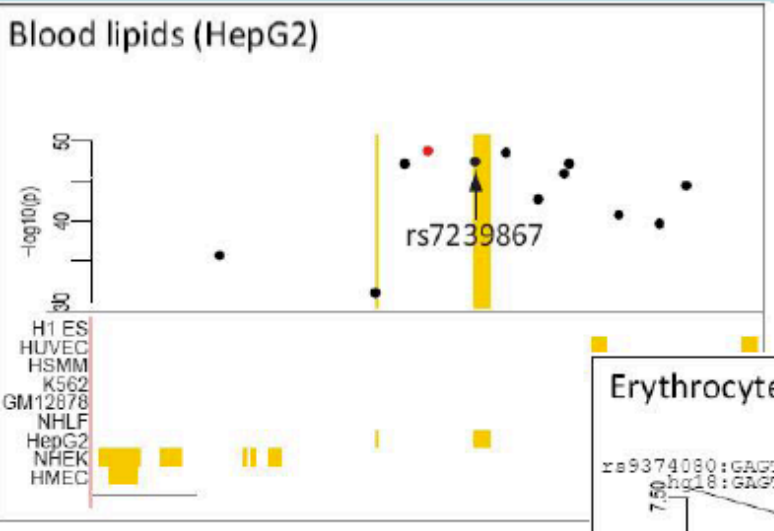
Disease SNPs correlate with enhancers that are active in related cell types

ENCODE

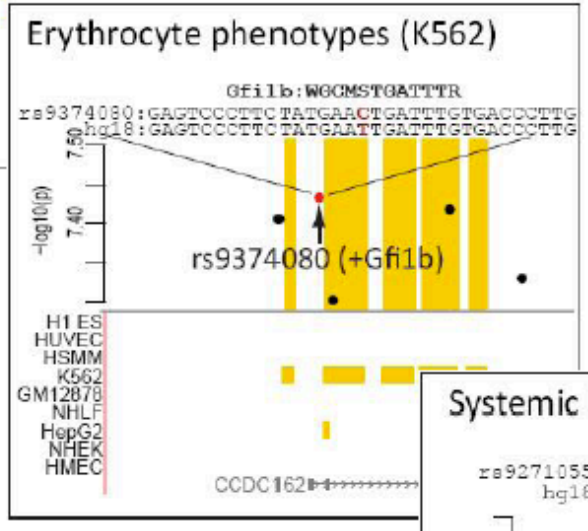
Phenotype	Top Cell Type	Total #SNPs from Study	#SNPs in enh. States 4 and 5	p-value	FDR	H1 ES	K562	GM12878	HepG2	HUVEC	HSMM	NHLF	NHEK	HMEC
Erythrocyte phenotypes (Ref. 38)	K562	35	9	$<10^{-7}$	0.02	9	17	4	0	0	1	2	1	1
Blood lipids (Ref. 39)	HepG2	101	13	$<10^{-7}$	0.02	3	5	0	11	2	3	3	4	3
Rheumatoid arthritis (Ref. 40)	GM12878	29	7	2.0×10^{-7}	0.03	0	0	13	0	2	0	0	2	3
Primary billary cirrhosis (Ref. 41)	GM12878	6	4	6.0×10^{-7}	0.03	0	11	41	0	0	0	0	8	8
Systemic lupus erythromatosus (Ref. 42)	GM12878	18	6	9.0×10^{-7}	0.03	0	4	21	0	5	8	0	3	5
Lipoprotein cholesterol/triglycerides (Ref. 43)	HepG2	18	5	1.2×10^{-6}	0.03	17	8	0	24	3	6	4	3	3
Hematological traits (Ref. 44)	K562	39	7	1.7×10^{-6}	0.03	0	12	10	2	1	0	0	1	0
Hematological parameters (Ref. 45)	K562	28	6	2.2×10^{-6}	0.03	0	15	7	0	5	7	7	3	2
Colorectal cancer (Ref. 46)	HepG2	4	3	3.8×10^{-6}	0.03	0	0	0	66	0	12	0	12	12
Blood pressure (Ref. 47)	K562	9	4	5.0×10^{-6}	0.04	0	30	14	0	10	6	7	5	11

Disease/phenotype		Enhancer specificity
Erythrocyte phenotypes	⟷	Erythrocytic cells
Lipids, cholesterol	⟷	Hepatic cells
Lupus, arthritis	⟷	Lymphoid cells

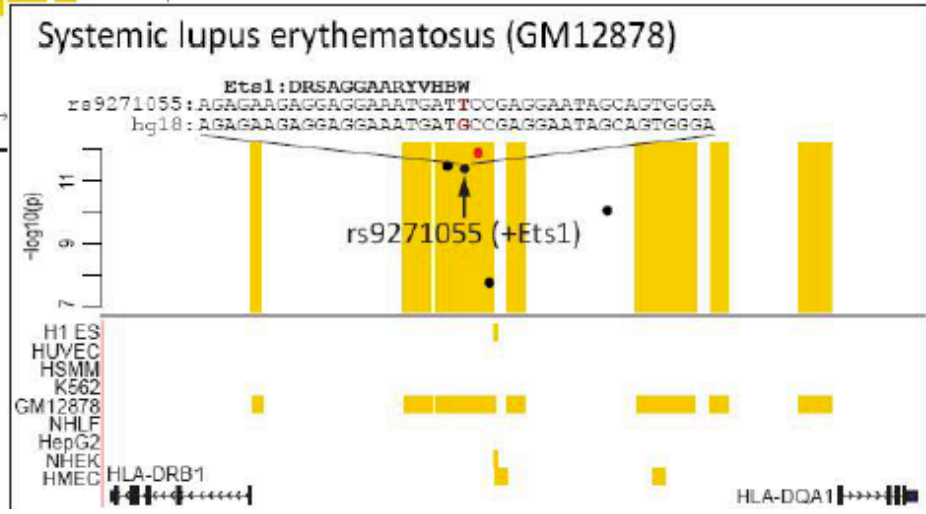
Annotations & regulatory predictions for GWAS



Associated SNP coincides with cell type-specific enhancer



SNP affects predicted regulatory interactions



Conservation of *RET* Regulatory Function from Human to Zebrafish Without Sequence Similarity

Shannon Fisher,[1,2]*† Elizabeth A. Grice,[1]* Ryan M. Vinton,[1] Seneca L. Bessling,[1] Andrew S. McCallion[1,3]†

Evolutionary sequence conservation is an accepted criterion to identify noncoding regulatory sequences. We have used a transposon-based transgenic assay in zebrafish to evaluate noncoding sequences at the zebrafish *ret* locus, conserved among teleosts, and at the human *RET* locus, conserved among mammals. Most teleost sequences directed *ret*-specific reporter gene expression, with many displaying overlapping regulatory control. The majority of human *RET* noncoding sequences also directed *ret*-specific expression in zebrafish. Thus, vast amounts of functional sequence information may exist that would not be detected by sequence similarity approaches.

A current hypothesis is that sequences conserved over greater evolutionary distances are more likely to be functional than those conserved over lesser distances (*1*). Many recent publications have focused attention on the regulatory potential of "ultra-conserved" noncoding sequences, conserved across great evolutionary distances, e.g., human to fugu (*2–9*) [≥300 million years, or average 74% protein identity (*10*)]. These are frequently enhancers associated with developmental genes, consistent with strong selective pressure to preserve critical mechanisms. Analyses of identified sequences have generally fallen into two categories: analyses confined to mammals, with functional verification done in mice, or analyses including mammalian and teleost sequences, focusing on highly conserved sequences alignable at the extremes. However, simply because an expression pattern is preserved through evolution, it does not necessarily follow that the cis-regulatory elements controlling that expression in one species will function in a second.

We have explicitly tested two hypotheses: First, using selective pressure as a guide across moderate evolutionary distances, we can identify the majority of enhancers control-

[1]McKusick–Nathans Institute of Genetic Medicine, [2]Department of Cell Biology, [3]Department of Comparative Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: sfisher@jhmi.edu (S.F.); amccalli@jhmi.edu (A.S.M.)

ling expressio
tional testing
manner, and
noncoding se
evolutionary
overt sequenc

We have f
of the gene en
kinase. *RET* i
genital precurs
during embry
and periphera
during develo
though *RET*
across evoluti
coding the tyr
conserved [≥7
humans to zeb
the genomic s
segment enco
with the ortho
using AVID/V
ZCS (zebrafish
corresponding
(table S1).

We also
conserved no
paring a ~2
human *RET*
intervals in
We selected s
and at least t
In total 13 H
amplicons, e
served seque
for analysis.

Copy Number (Structural) Variation

Structural variation of the genome

kilobase- to megabase-sized

deletions,

duplications,

insertions,

Inversions

complex combinations of rearrangements.

CNVs due do non-allelic recombination between low-copy repeats (LCRs) that lead to human disease

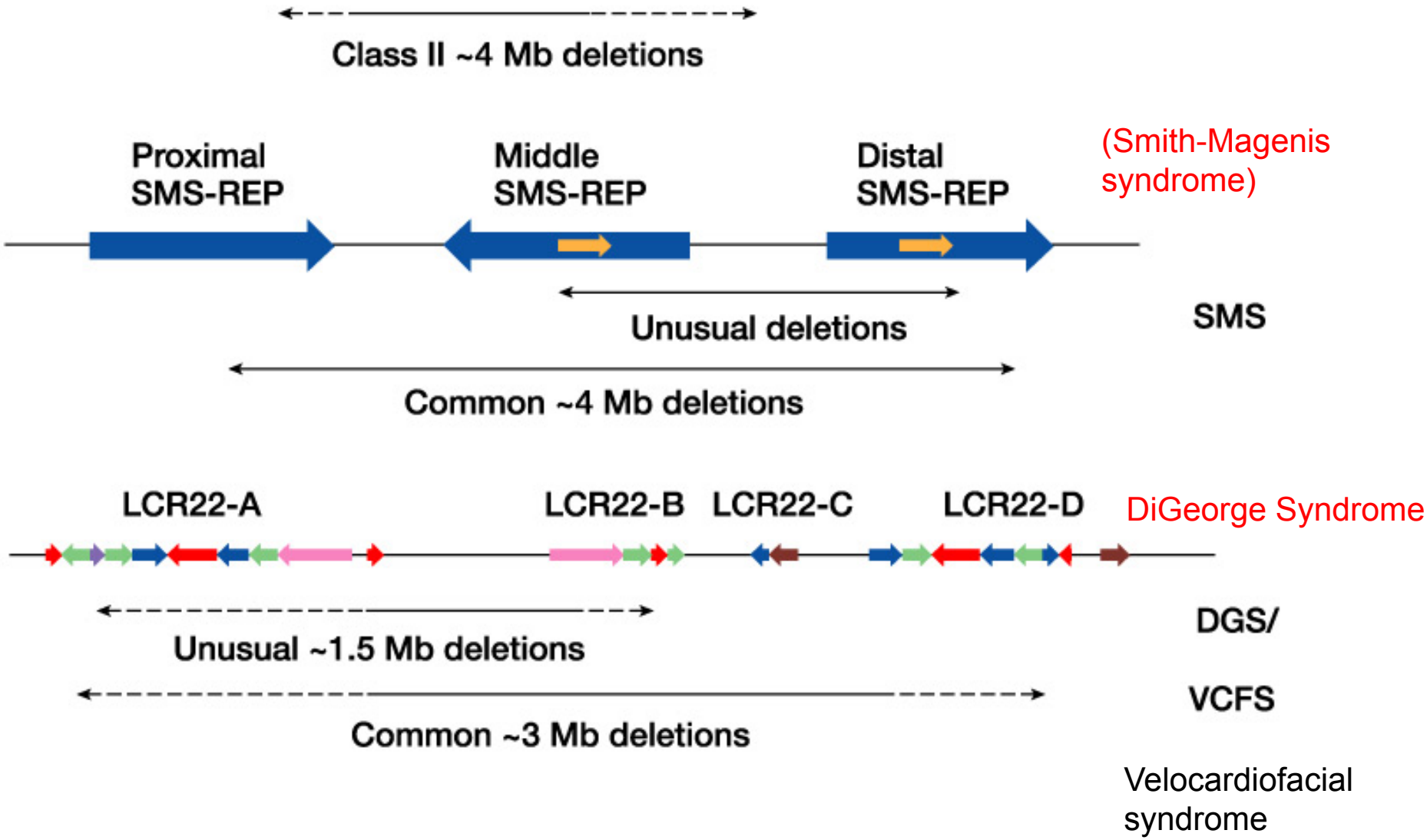


Figure 11-19 part 2 of 2 Human Molecular Genetics, 3/e. (© Garland Science 2004)

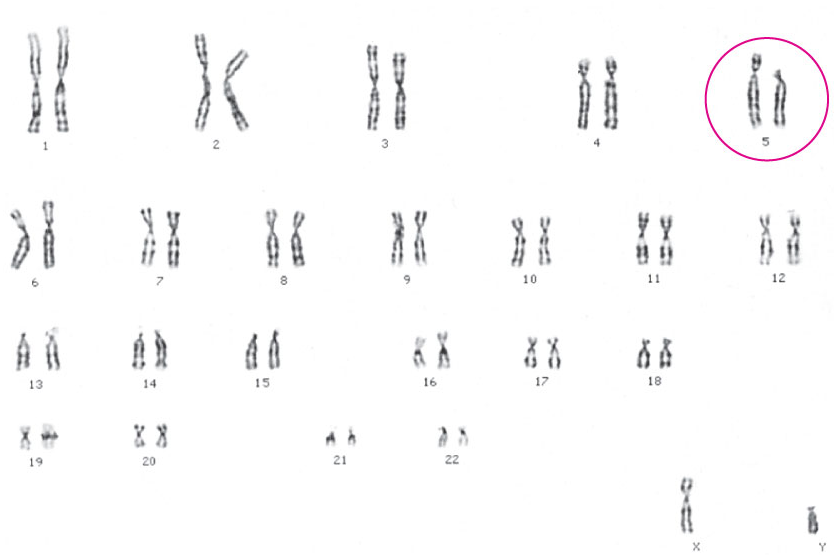
Examples of contiguous gene syndromes

- Xp:
  - Successively large deletions remove more genes and add more diseases
- 11p12:
  - WAGR (Wilm's tumor, Aniridia, Genital and/or urinary tract abnormalities, Mental retardation)

Segmental aneuploidy

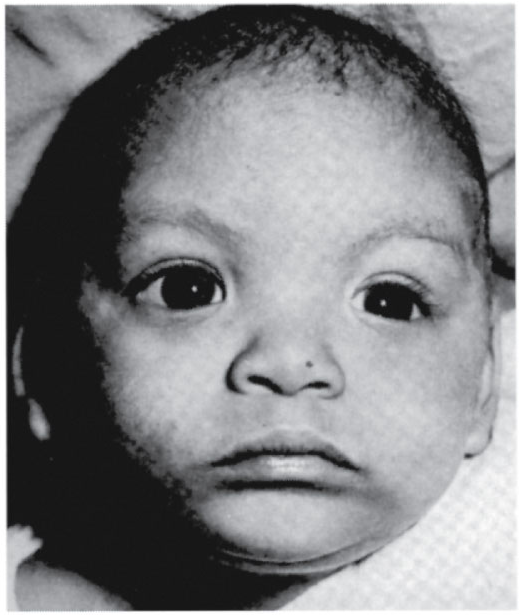
- Phenotype in heterozygotes depends on only a subset of deleted genes that are dosage sensitive
- De Novo microdeletions, frequently flanked by long repeats (often transcribed hence open chromatin - more recombination prone)

a) Karyotype (G banding)



46, 5p-

b) Individual with Cri-du-chat syndrome



Adapted from Figure © 2010 PJ Russell, iGenetics 3rd ed.

Williams syndrome - example of segmental aneuploidy (1.6Mb deletion at 7q11.23/~ 20 genes)

1/20,000 births

Growth retardation

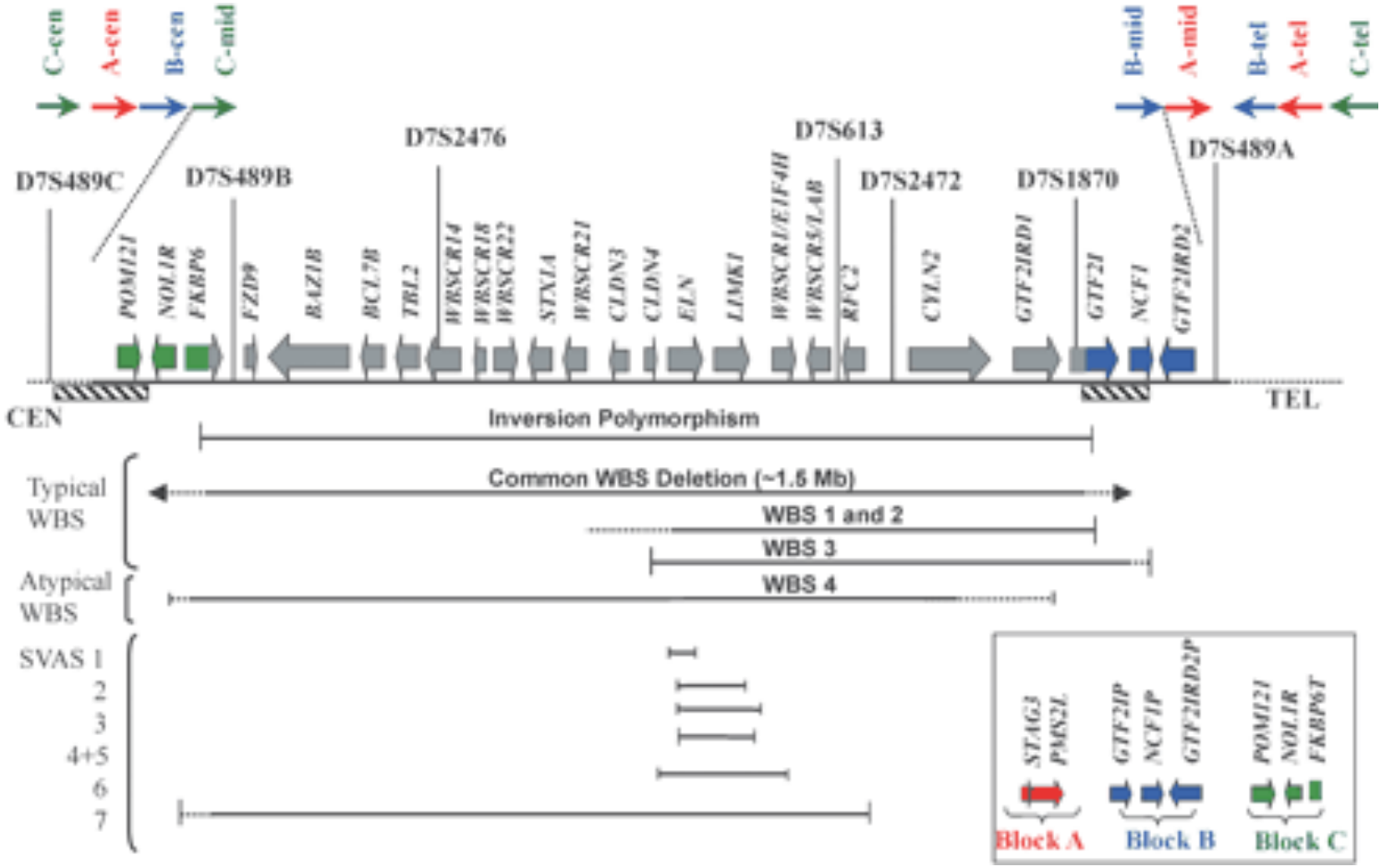
Hypercalcemia

Sypervalvular aortic stenosis (elastin)

Moderately mentally retarded

Highly sociable, often musical, defect in visuospatial constructive ability

Deletions identified in Williams syndrome



Current knowledge of of CNV (copy number variation) in the human genome

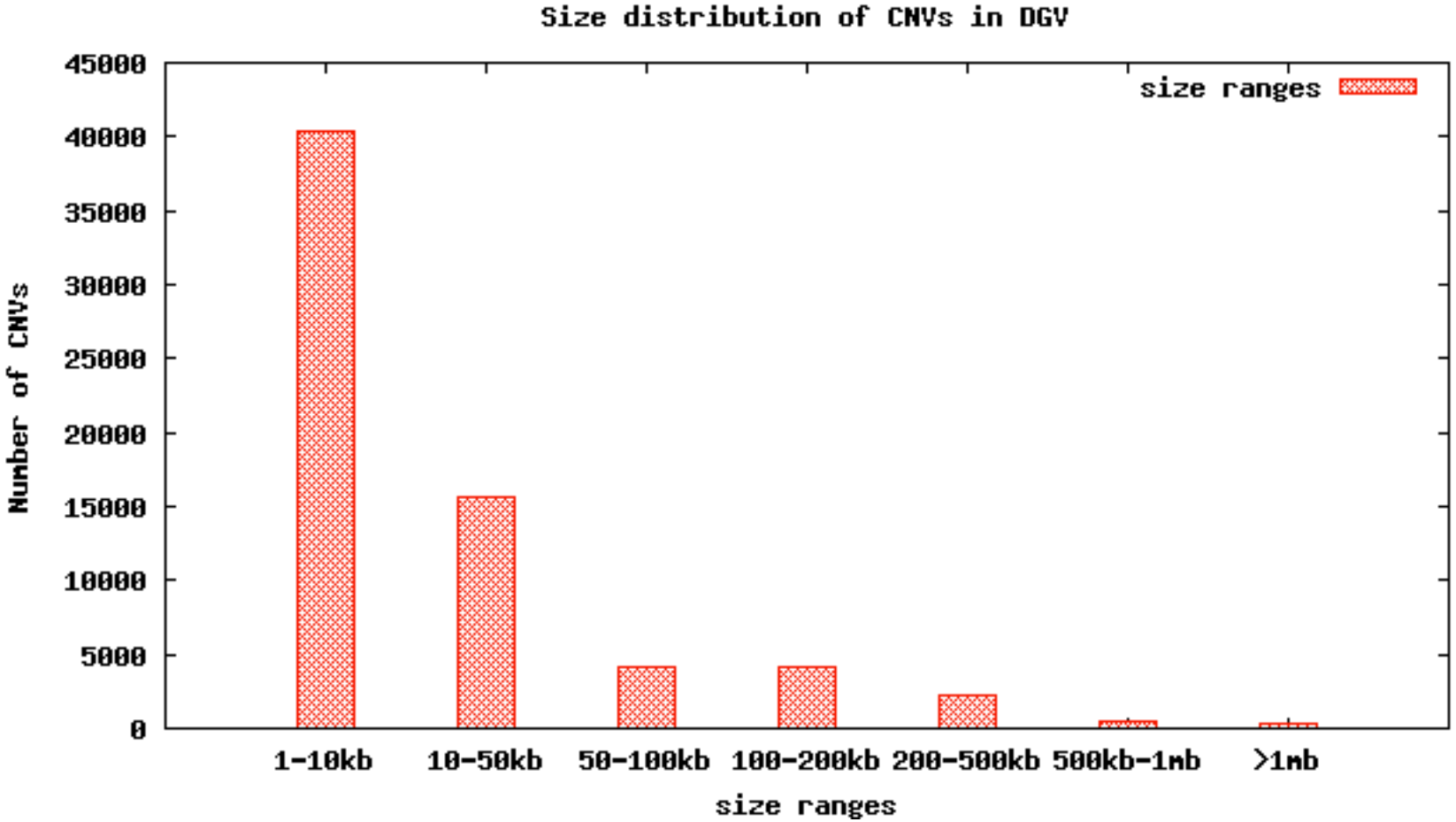
Database of Genomic Variants

http://projects.tcag.ca/variation/

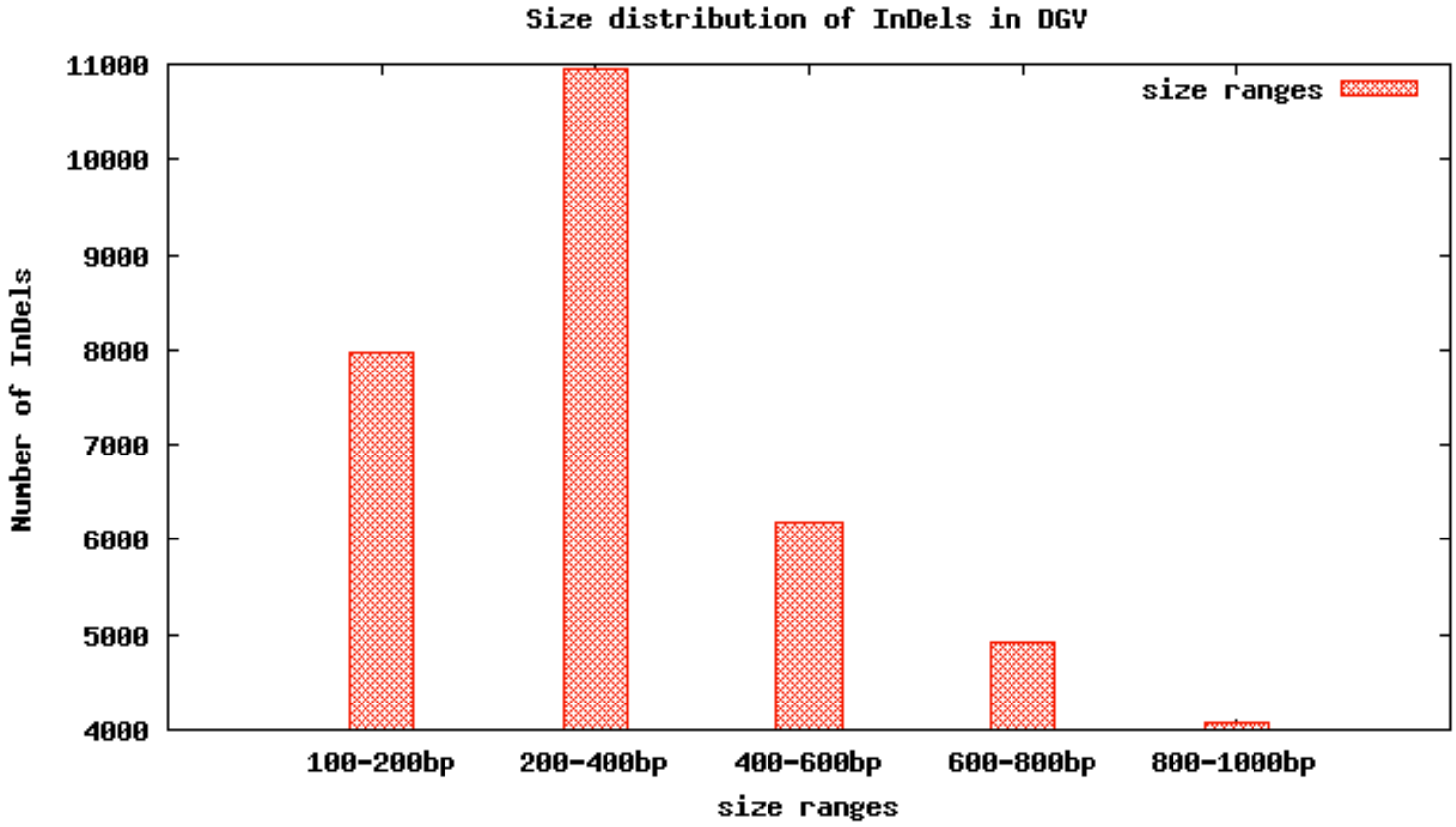
011/10: Total entries: 101,923 (hg18)

- CNVs: 66,741
- Inversions: 953
- InDels (100bp-1Kb): 34,229
- Total CNV loci: 15,963
- Articles cited: 42

Size distribution (CNV):
The graph displays the size distribution of CNVs in the database.



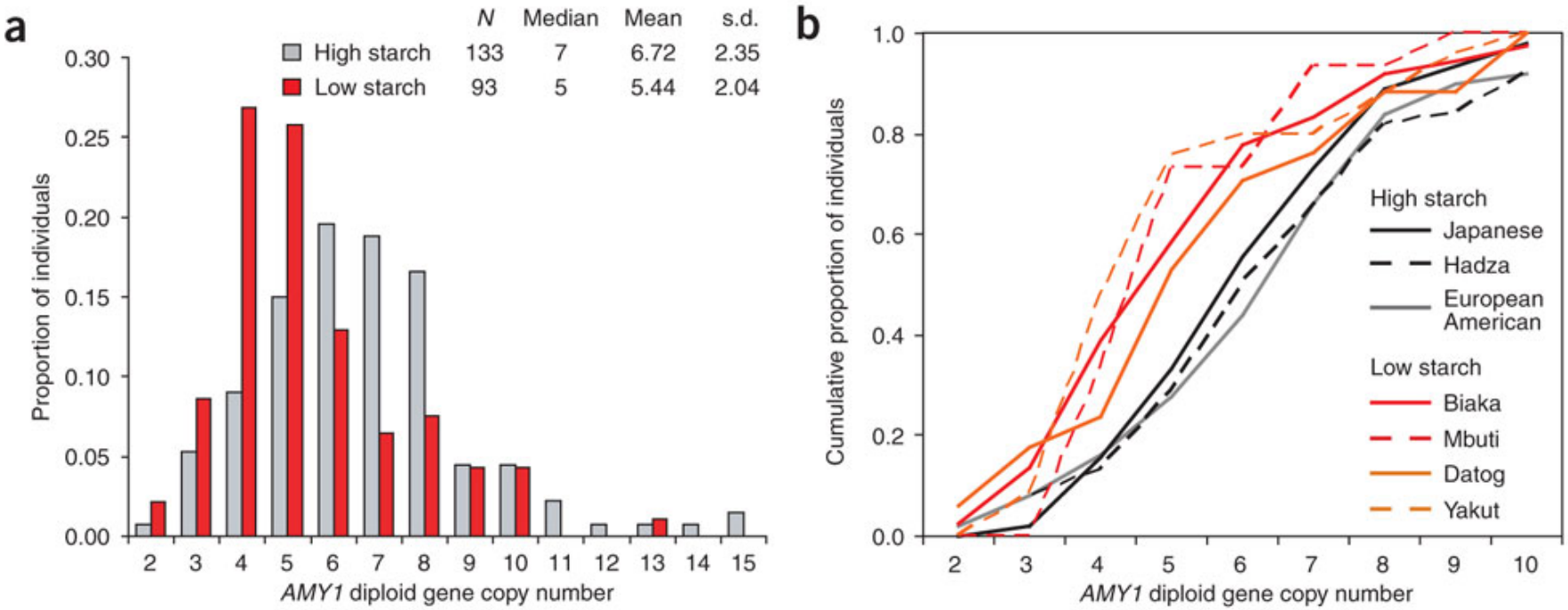
Size distribution (InDel):
The distribution of InDel sizes (100bp-1kb) are shown here. The noticeable increase of variants in the ~300bp size interval likely reflects the abundance of polymorphic ALU repeats in the human genome.



CNV affecting one human phenotype: Ability to digest starch:

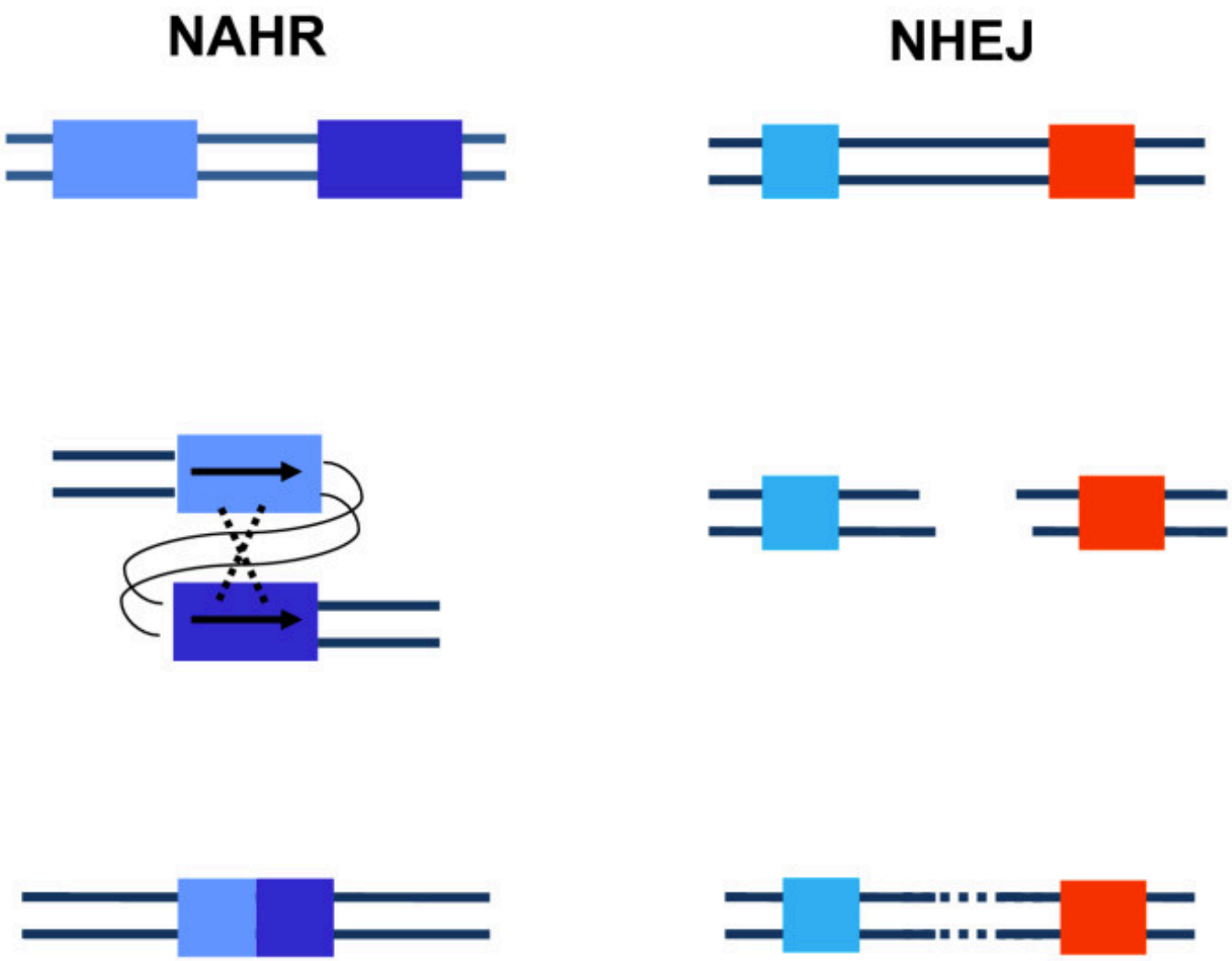
Diet and the evolution of human amylase gene copy number variation

Perry et al. Nature Genet. Nature Genetics 39, 1256 – 1260 (2007)



Different types of mechanisms leading to SVs/CNVs

- Non-allelic homologous recombination (NAHR)
- Non homologous end joining (NHEJ)
- Transposon insertion (e.g. L1)



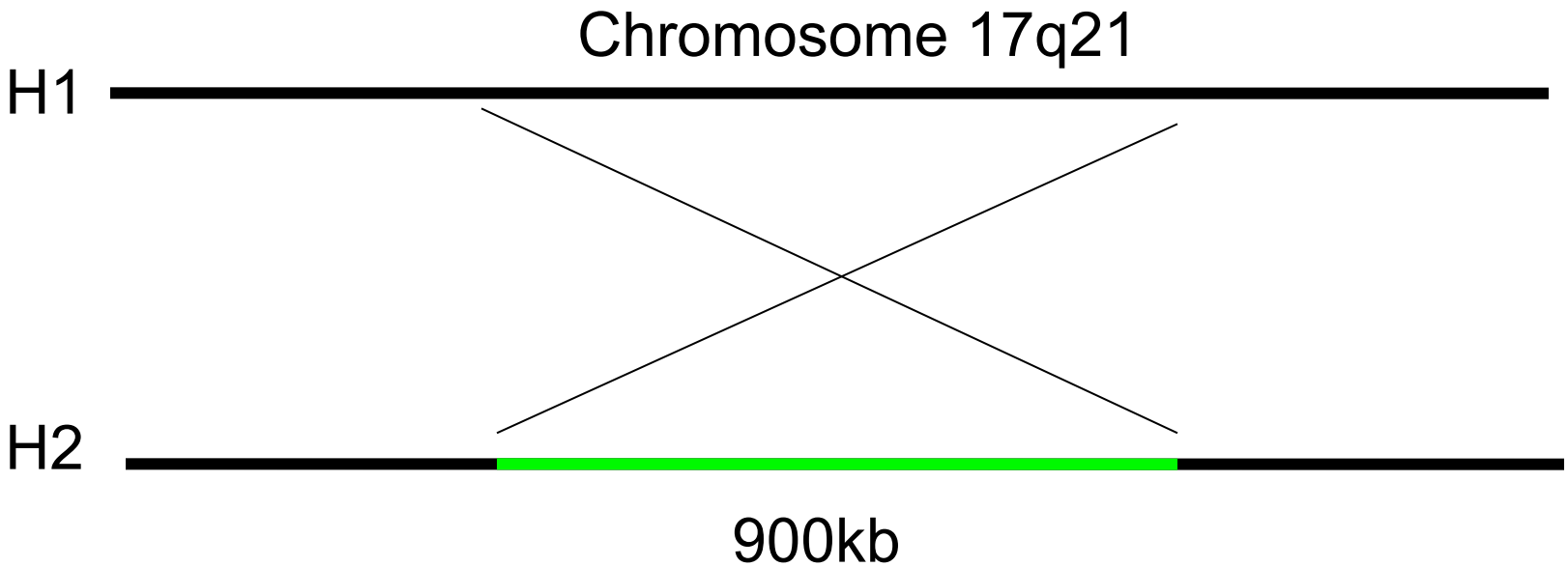
Gu *et al. PathoGenetics* 2008 1:4

Genomic rearrangements in the genome are likely to be more common than expected

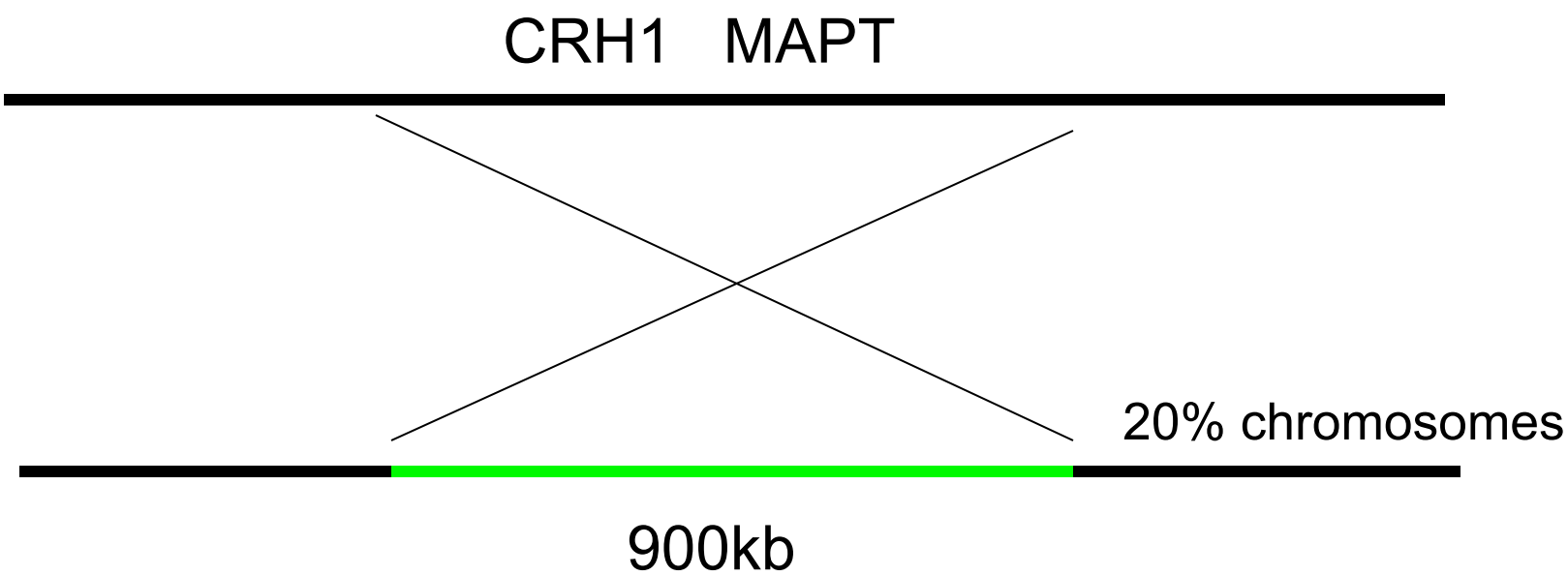
5% of human genes are found in interspersed duplicated copies

Recent examples of large rearrangement polymorphisms in the human genome

- Sebat et al. **Large-scale copy number polymorphism in the human genome.**
  Science 305:525-8 (2004)
- **A common inversion under selection in Europeans**
  Stefansson et al. *Nature Genetics* **37**, 129 - 137 (2005)

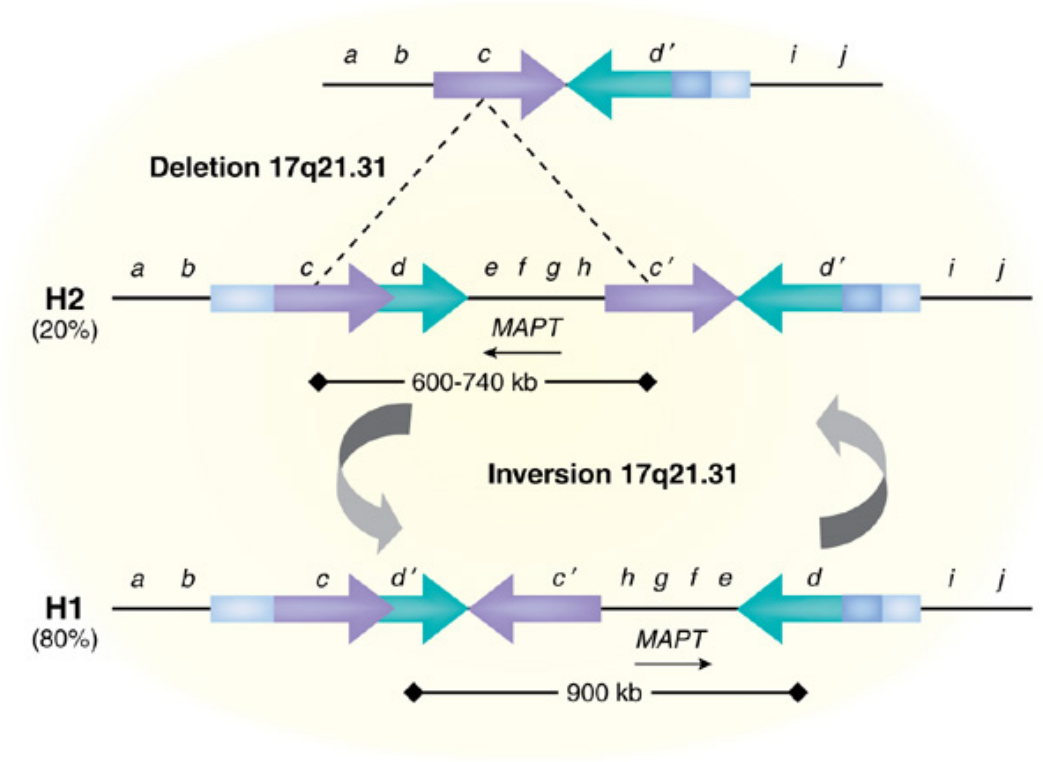


A common inversion under selection in Europeans
Stefansson et al. *Nature Genetics* **37**, 129 - 137 (2005)



CRH1: corticotropin releasing hormone receptor 1
MAPT: microtubule associated protein tau

Genomic architecture, rearrangements and marker genotypes at 17q21.31



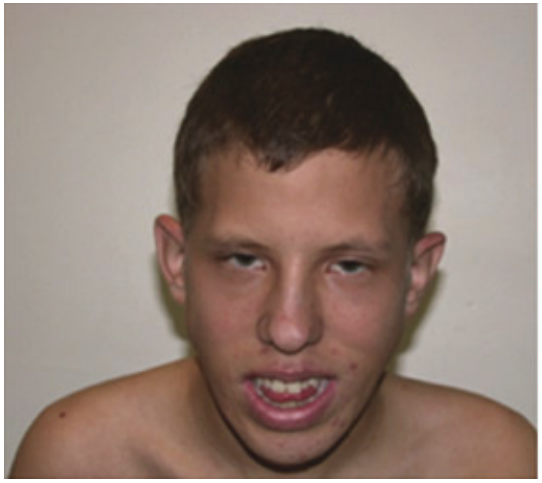
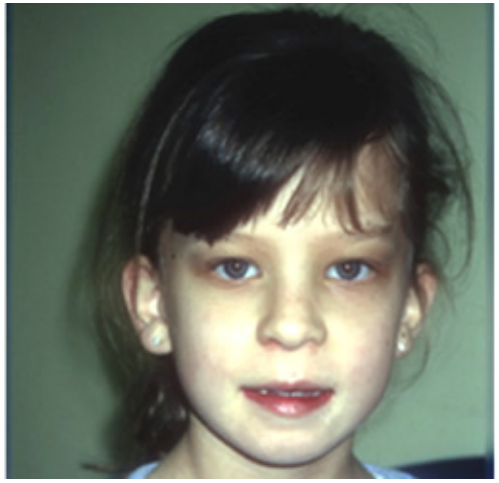
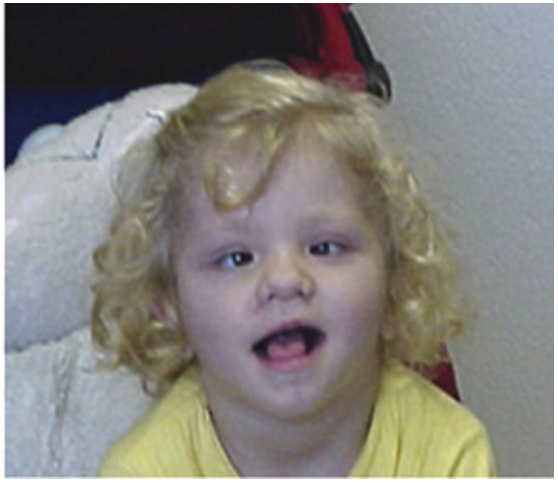
Microdeletion syndromes:

1. Koolen, D.A. *et al*. *Nat. Genet.* **38**, 999–1001 (2006).

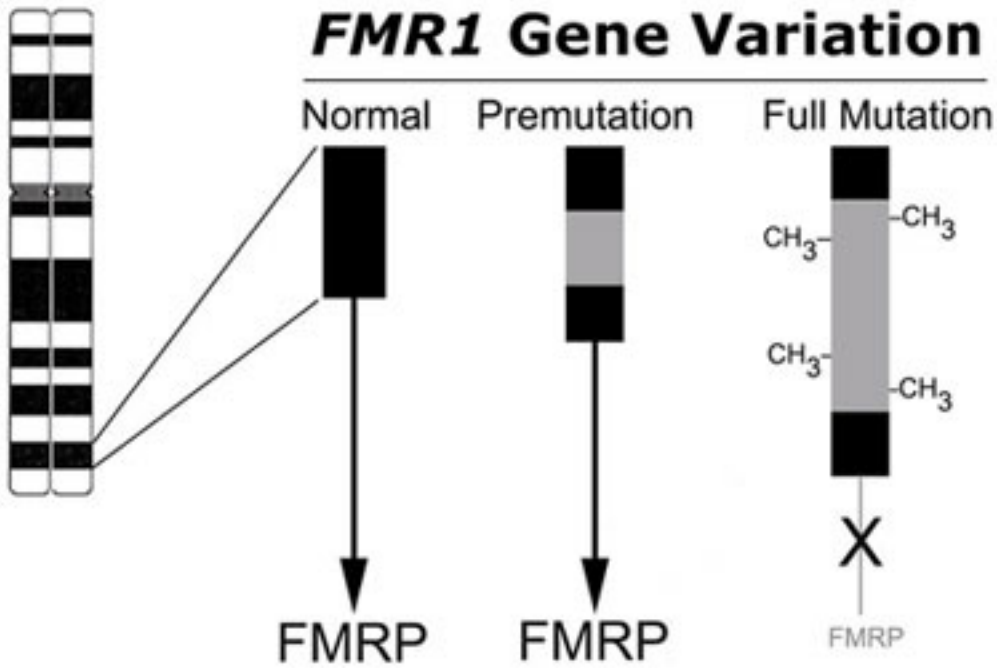
2. Shaw-Smith, C. *et al*. *Nat. Genet.* **38**, 1032–1037 (2006).

3. Sharp, A.J. *et al*. *Nat. Genet.* **38**, 1038–1042 (2006).

Affected individuals with 17q21 microdeletion (1% of mental retardation patients)

Fragile X – trinucleotide repeat



- Most common inherited form of mental retardation
- Due to instable CGG repeat at FMR1
- All full mutations derive from premutation (56-200 repeats)
- Expansion through female meiosis
- Severity correlates with CGG repeats

Rare L1 insertions as a cause of disease

LINEs (long interspersed repetitive elements)

AKA: Kpn1 element

~5kb. Relic of retrovirus

3 distantly related LINE families are found in the human genome, but only LINE1 is active.

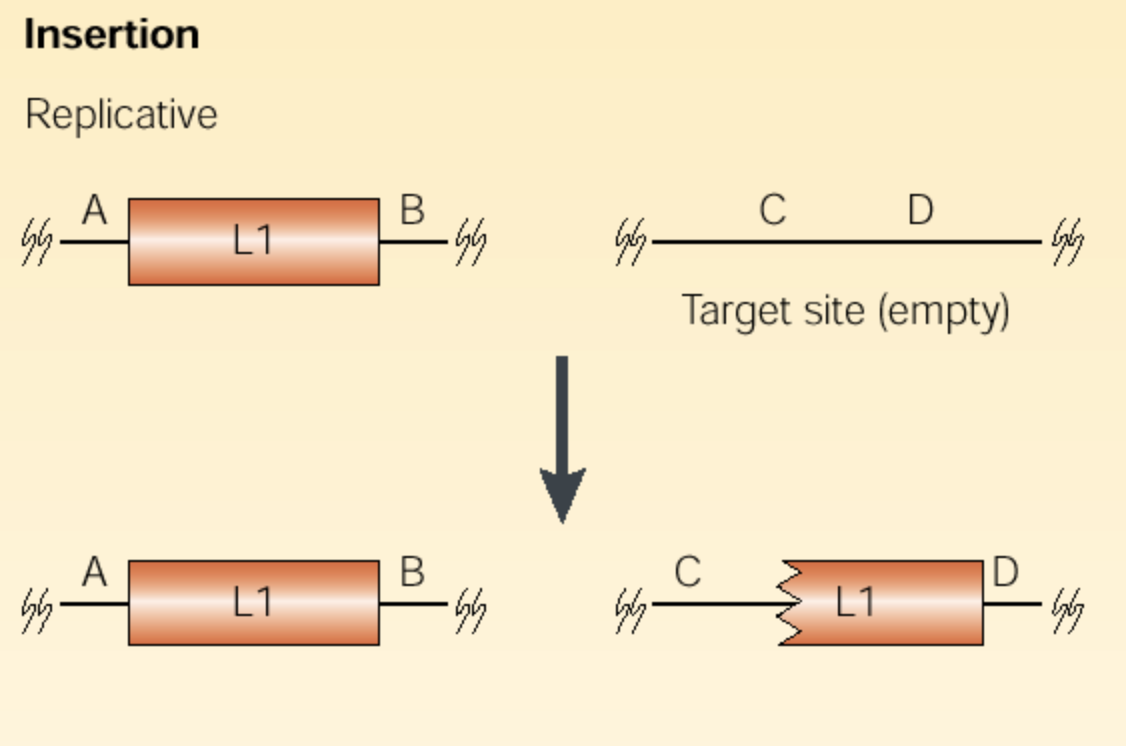
Human genome: ~515,000 copies of LINE1 (L1), ~365,000 L2, and ~37,000 L3 (most are truncated or rearranged)

Only ~30-60 are active

In mouse, ~3,000 are active.

L1 Insertions

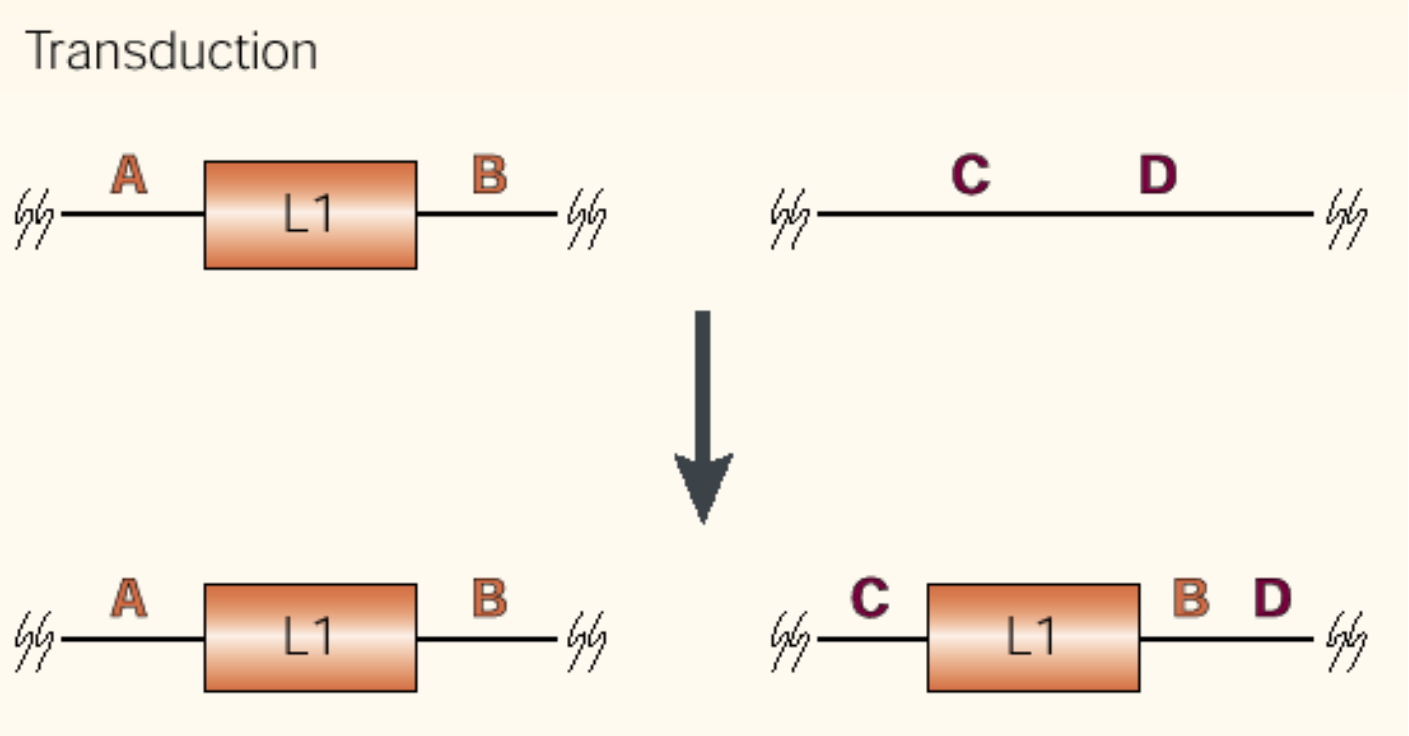
Occasionally lead to genetic disorder: Reported in Duchenne muscular dystrophy, type 2 retinitis pigmentosa, thalassaemia, chronic granulomatous disease, and hemophilia A (2/140 patients) - insertion into factor VIII.



Prak & Kazazian. (2000) NRG. 1: 134-144

L1 elements also cause Transduction.

In addition to duplicating themselves, L1s can carry with them genomic flanking sequences that are downstream of their 3UTRs.



Prak & Kazazian. (2000) NRG. 1: 134-144

Mitochondrial genetics

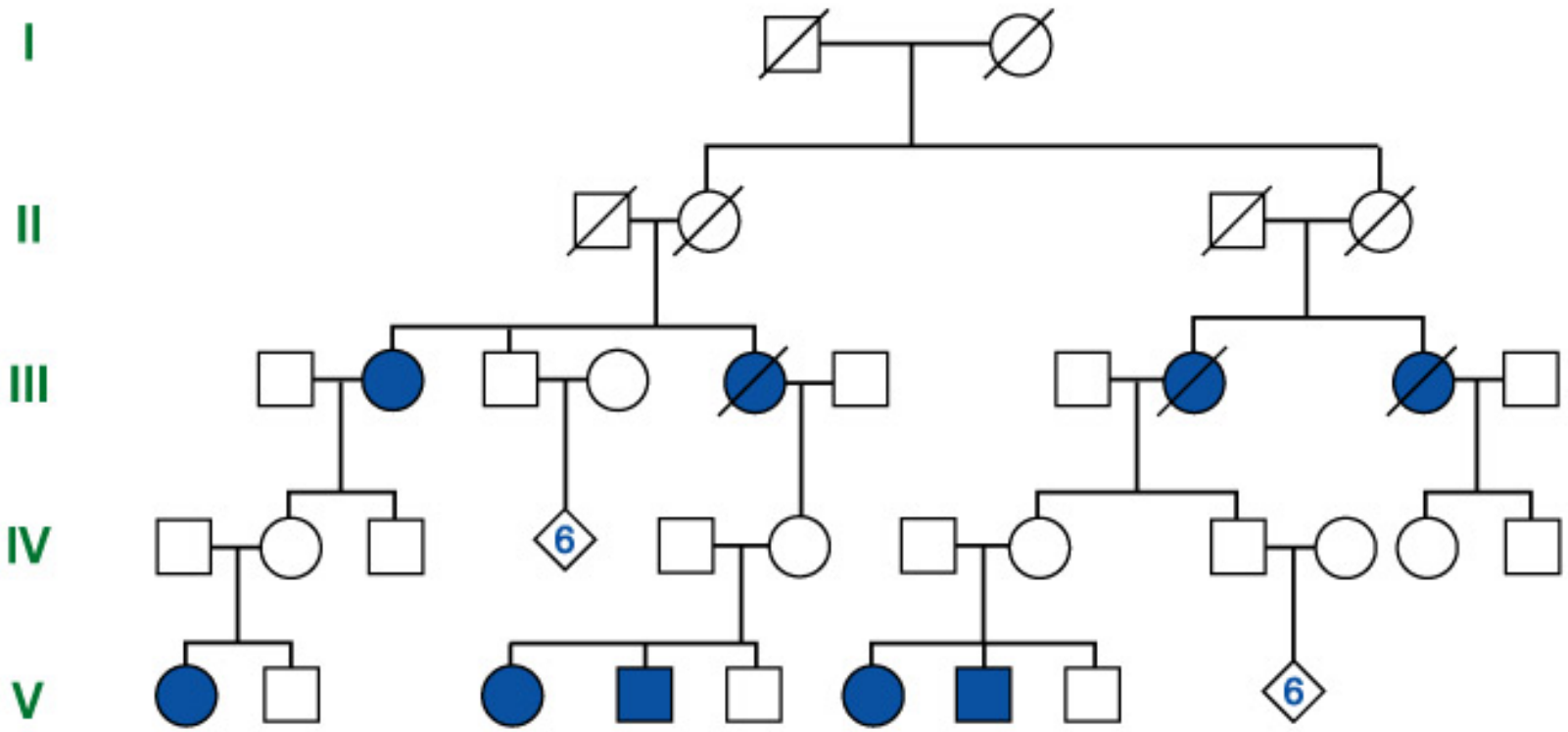
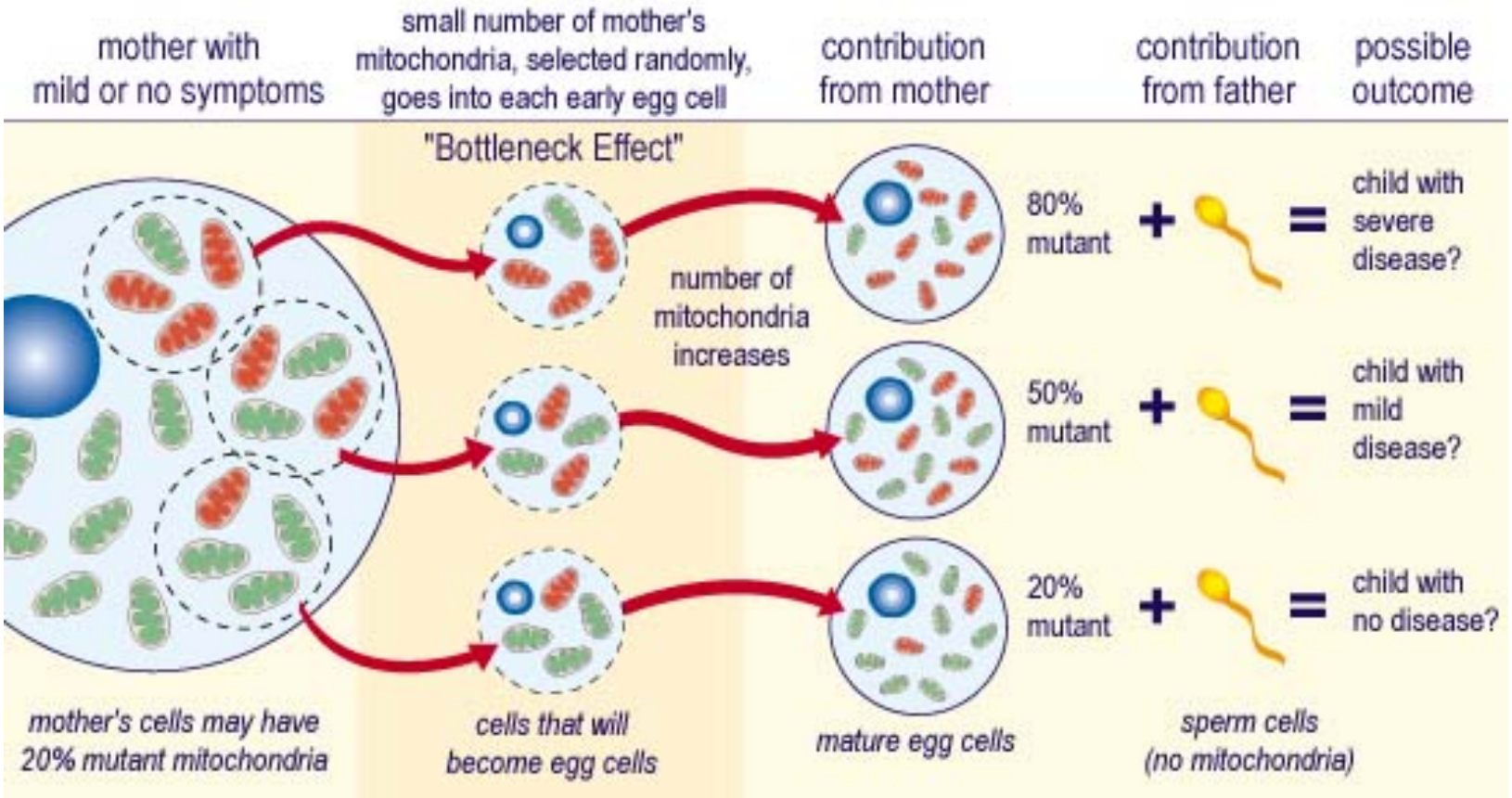


Figure 4-4 Human Molecular Genetics, 3/e. (© Garland Science 2004)

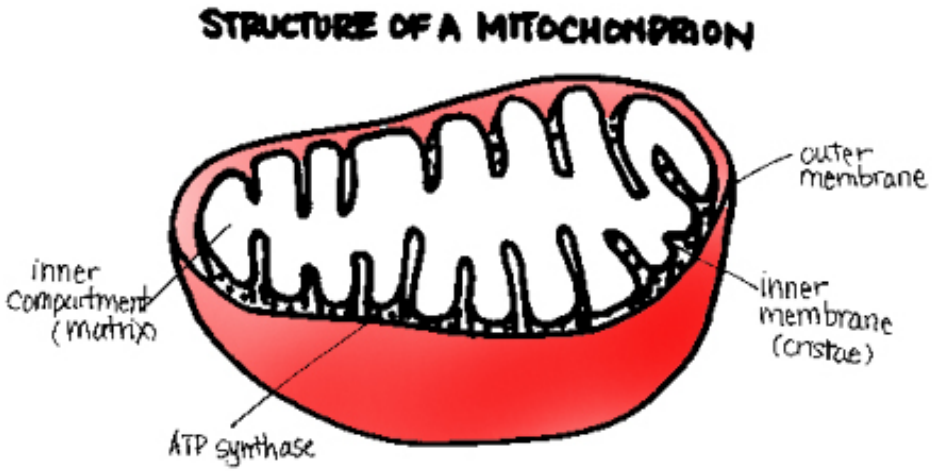
Mitochondrial inheritance gives matrilineal pedigree pattern

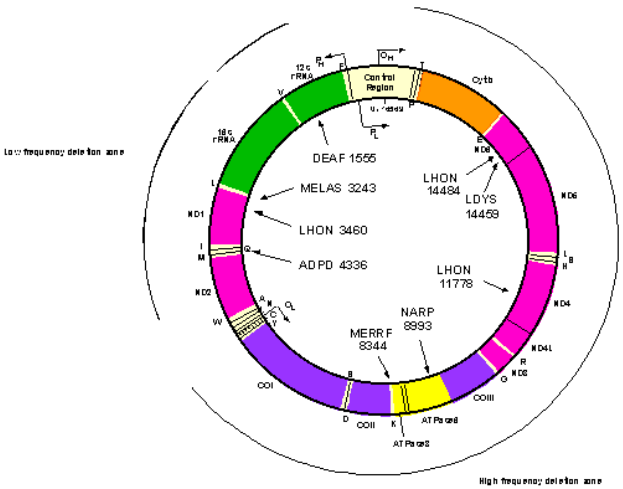
- Matrilineal inheritance
- Sperm mtDNA is actively degraded
- Mitochondria present in thousands of copy per somatic cell
- Normal individuals: ~99.9% of molecules are identical (homoplasmy)
- New mutation leads to heteroplasmy
- In some patients with mitochondrial disease every patient carries causative mutation (homoplasmy)
- In some patients there is heteroplasmy.

The mitochondrial genetic bottleneck



Mitochondrial genome





- Cytoplasmic organelle
- Small genome (~16kb)
- 1/200,000 size of nuclear genome
- Sequenced in 1981 (Anderson et al.)
- 93% is coding
- ~ 1,000 copies per cell

Organ systems and symptoms of mt diseases

- Multisystem disorders with large range of symptoms:
- Brain, heart, skeletal muscle, kidney and endocrine systems can be affected (sometimes there is a threshold effect)
- Symptoms: forms of blindness, deafness, movement disorders, dementias, heart disease, muscle weakness, kidney dysfunction, endocrine disorders (including diabetes)

Functions of the 37 mitochondrial genes

- 13 of mitochondrial peptide subunits in mitochondrial respiratory-chain complex (OXPHOS)
  - Remaining > 67 OXPHOS subunits are nuclear encoded
- rRNAs: 2
- tRNAs: 22; Located between every 2 rRNA or Protein coding genes
- Third base wobble
- All factors involved in maintenance,replication & expression of mtDNA are Nuclear encoded

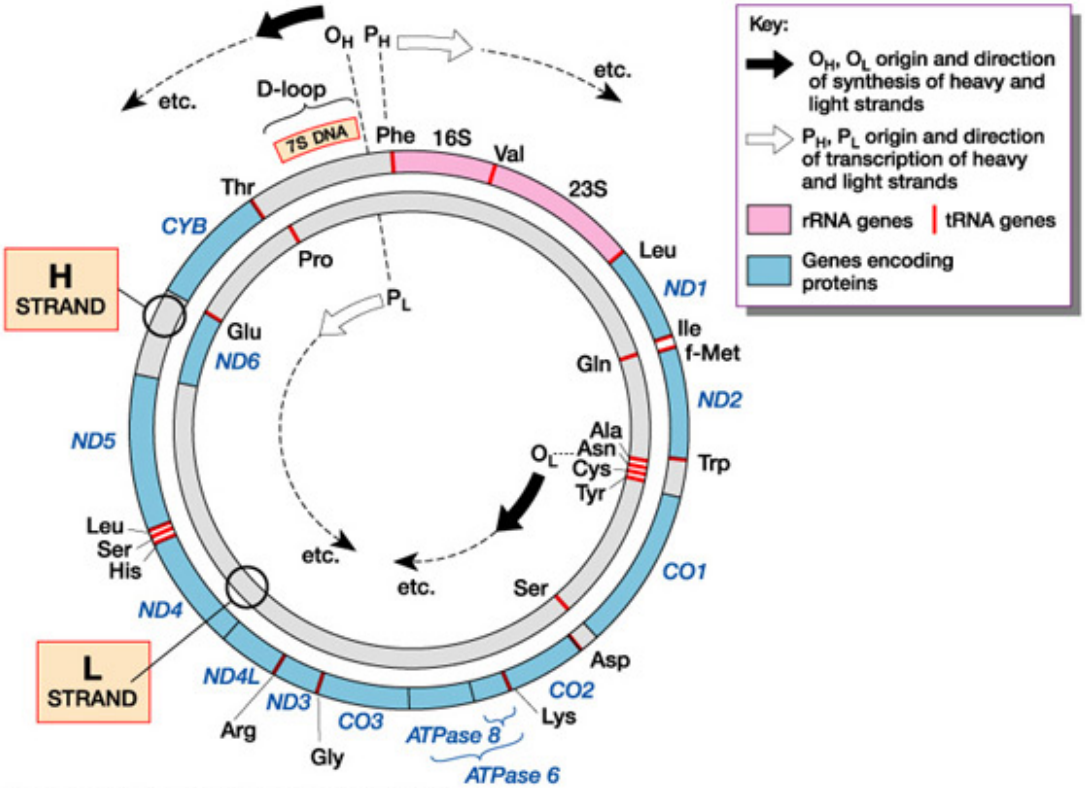


Figure 9-2 Human Molecular Genetics, 3/e. (© Garland Science 2004)

Example of moderately severe tRNA mutations

tRNA Lys :

- A8344G (Frequent)
- T8356C
- G8363A
- G8361A

MERRF (Myoclonic Epilepsy and Ragged-Red Fiber Disease)

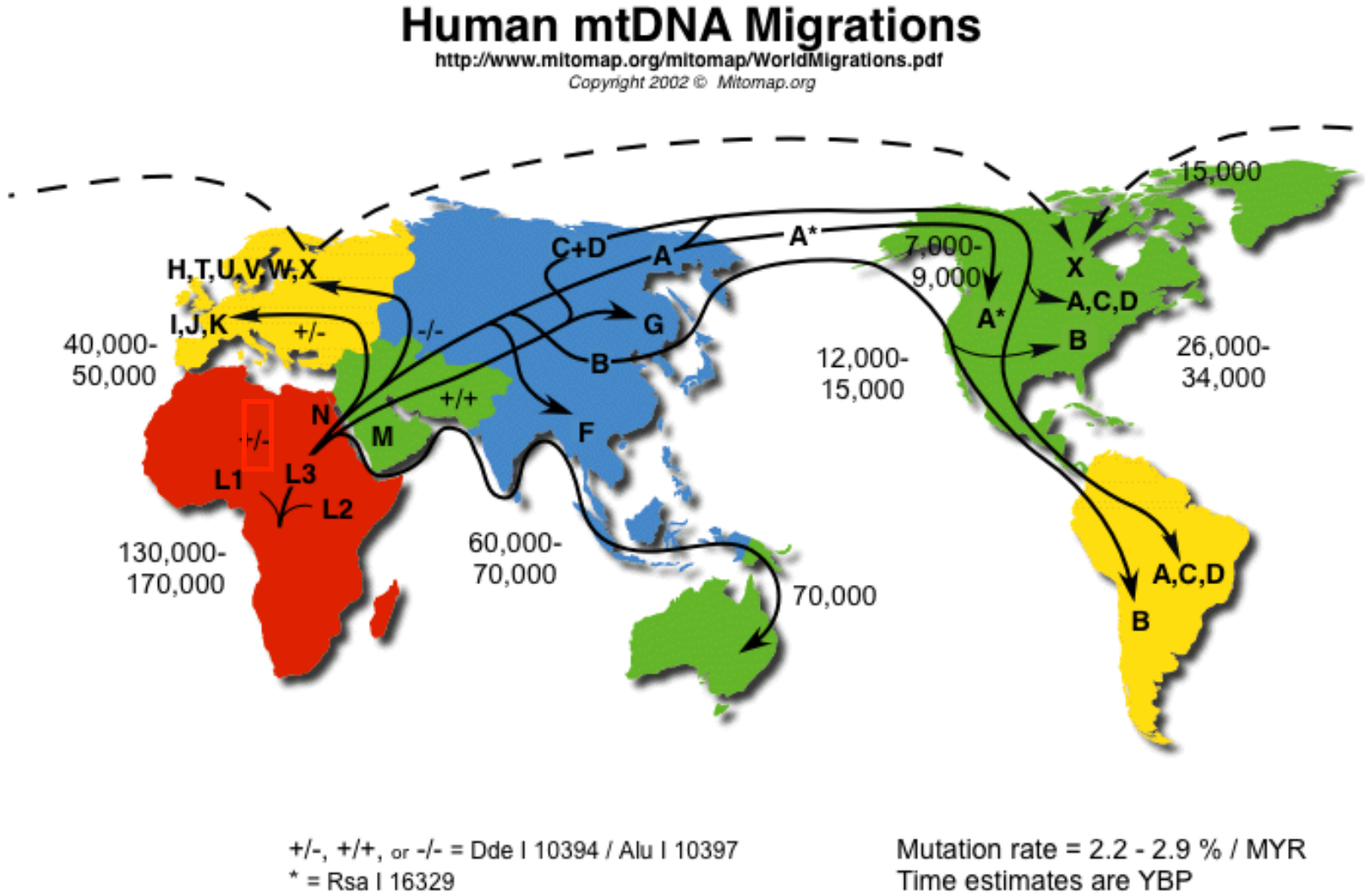
Onset: Late adolescence - Early adult

Level of mutant heteroplasmy + age of patient influence severity of symptoms

Mitochondrial DNA and human evolution: *Nature* **325**, 31 – 36 (1987)

Rebecca L. Cann*, Mark Stoneking & Allan C. Wilson

Mitochondrial DNAs from 147 people, drawn from five geographic populations have been analyzed by restriction mapping. All these mitochondrial DNAs stem from one woman who is postulated to have lived about 200,000 years ago, probably in Africa. All the populations examined except the African population have multiple origins, implying that each area was colonized repeatedly



J lineage often seen with Leber's hereditary optic neuropathy (LHON) patients

Identifying Human Variation

LETTER

doi:10.1038/nature09774

Human-specific loss of regulatory DNA and the evolution of human-specific traits

Cory Y. McLean[1]*, Philip L. Reno[2,3]*†, Alex A. Pollen[2]*, Abraham I. Bassan[2], Terence D. Capellini[2], Catherine Guenther[2,3], Vahan B. Indjeian[2,3], Xinhong Lim[2], Douglas B. Menke[2,3]†, Bruce T. Schaar[2], Aaron M. Wenger[1], Gill Bejerano[1,2] & David M. Kingsley[2,3]

hCONDEL – loss of enhancer for 1) sensory vibrissae, 2) penile spine enhancer in androgen receptor, loss of enhancer in tumor suppressor gene

Vol 443|14 September 2006|doi:10.1038/nature05113 nature

ARTICLES

An RNA gene expressed during cortical development evolved rapidly in humans

Katherine S. Pollard[1*†], Sofie R. Salama[1,2*], Nelle Lambert[4,5], Marie-Alexandra Lambot[4], Sandra Coppens[4], Jakob S. Pedersen[1], Sol Katzman[1], Bryan King[1,2], Courtney Onodera[1], Adam Siepel[1†], Andrew D. Kern[1], Colette Dehay[6,7], Haller Igel[3], Manuel Ares Jr[3], Pierre Vanderhaeghen[4] & David Haussler[1,2]

HAR1 – part of noncoding RNA expressed in human brain that has undergone expansion

HAR2 – enhancer involved in opposing thumb

February 2011

RESEARCH PERSPECTIVE