Bi 188 2013

April 5, 2013

Class 024 Kerckhoff, 3:00-5:00pm Fridays

BI 188 Human Genetics and Genomics

Meeting time: Fridays 3:00-4:55 in 024 Kerckhoff

General Notes:

Text: *Recombinant DNA: Genes and Genomes – A Short Course*, 3rd edition 2007
Authors: J. Watson, A. Caudy, R. Myers, and J. Witkowski
ISBN: 0-7167-2866-4

Course Website: http://woldlab.caltech.edu/bi188/

T.A.s Katherine Fisher-Aylor **kfisher@caltech.edu**
Say-Tar Goh **sgoh@caltech.edu**
Georgi Marinov **georgi@caltech.edu**

1. The book supplements the lectures, but it does not contain them. The book is intended to be
   > background material
   > chapters 1-7 are for filling in and brushing up on relevant molecular biology.

2. Most lectures will have some additional reading from the literature. Generally, this will include one or two review or summary pieces (which are best to read first) and one research paper. These accompany the lectures. You will download them using web of science etc.

3. There will be a midterm, a final exam, and problem exercises of two types: computational and “conventional”: Extra points > 100 are offered; deploy as suits you

Bi188 website: http://woldlab.caltech.edu/bi188/index.shtml
Username: student
Password: MudNoud6

Exam plan; Problem sets; Computational Problems

Katherine Fisher-Aylor	x4923	kfisher@caltech.edu
Georgi Marinov	x4923	georgi@caltech.edu
Say-Tar Goh	x4923	sgoh@caltech.edu

Office hours – 128 Kerckhoff   finalized at first class meeting

30 points midterm, closed notes and other resources, out 5/3 3:00pm; Due 5/7 3pm
40 points final
3 computational genomics exercises 10 points each
4 non-computational problem sets 7.5 points each

Course scored for final grade based on 100 point max scale; points accrued above 100 can add a + to an A.

Unless altered by circumstance (ie Ditch Day) problems are due Friday, Beginning of class (3:00pm) electronically.

First set due Fri 3:00 pm April 12.

Computational tutorials and 3 computational analysis exercises

1. Map sequence read data for exomes,
   call candidate mutations and analyze: out 4 8/9; due 4/19
   Supporting python tutorial 4/8 and 4/9
   Supporting analysis discussion in class 4/12

2. RNA-seq data as FPKM; classify tumor types
   Out 4/26 due 5/10 3pm
   Supporting class presentation 4/26
   Supporting discussion 5/3 (after review for midterm)

3. Tumor/ normal genome comparison: diagnose the case, suggest action
   integrate RNA, DNA, methylation (as tracks)
   Out 5-17; due 5-31 In class discussions 5-17 and 5-24

Human Genetics and Genomics: Multiple Scientific and Societal Goals

I. Basic Biology Discovery – Use mutation / variation to identify a process; figure out its protein and RNA components -> clues to mechanism of action

A. classical or “forward” genetics = begin with a mutation; find the gene; study mutated individuals
essence: start with a trait or phenotype and work toward causal gene(s)

B. “reverse genetics” = you know the gene; mutate it in a model organism, cells, or find the mutations in humans by DNA screening
essence: start with a DNA variation (or gene) and work toward phenotype

II. Medical Genetics and Genomics. Infinite Expectations Where do we stand?

A. Better diagnosis of disease: genetic contribution. "Precision Medicine"

Cancer is prominent disease of genome and epigenome

Germline (BRCA1,2; TP53, Rb & other known and unknown)

Somatic (Hundreds of genes – pathway synthesis)

Single gene traits (Cystic Fibrosis; Muscular Dystrophies; Globinopathies)

Complex multigenic traits (i.e. Diabetes type 2; autism)

Chromosome level variations (i.e. Downs etc)

B. New and better treatment of disease

Gene therapy (conceptually beautiful; slow and difficult to bring to fruition; yet positive examples now coming on) Prof. Hacia last lecture

Make novel drugs

small molecule screens - Gleevec etc

therapeutic antibodies – Herceptin etc

other – at extreme, complete one-off custom solutions

C. Future for science and society: Will the Genome Information Commons become a reality? ELSI issues. Cost and delivery challenges under current models in US/ elsewhere.

Sequencing big eukaryotic genomes: How it was done & how DNA sequencing has changed since

Human was project impetus – “completed” 2003 (draft 2001)

2 projects A. The clone-based hierarchical shotgun by public consortium

- Multiple individual genomes in the aggregate assembly; one individual per BAC region

Subsequent “ finishing” to <10^{-4} error rate

Some areas remain unfinished (centromeres, telomeres, and 357 gaps in Build HG19).

Primary Reference paper:XXX

Focused research review on structural variation:XXX

Pertinent science history: http://dx.doi.org/10.1016/S0022-2836(02)00333-9

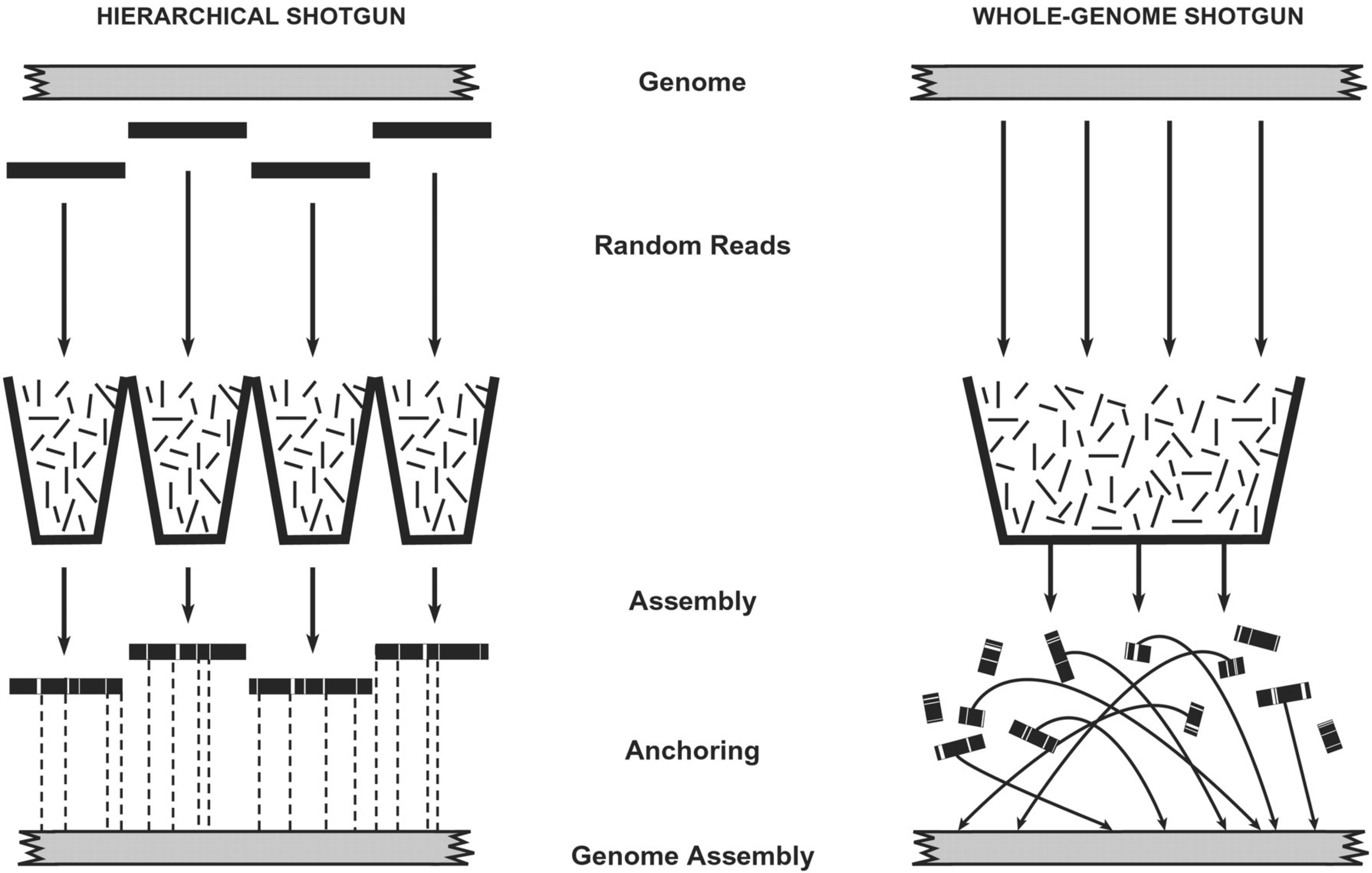
B. Second was the first mammalian whole-genome shotgun assembly (WGS) done by Celera Inc. Now this is of largely historic interest

- no finishing was done in the Celera project; they incorporated public project data
- one individual’s genome (Craig Venter)

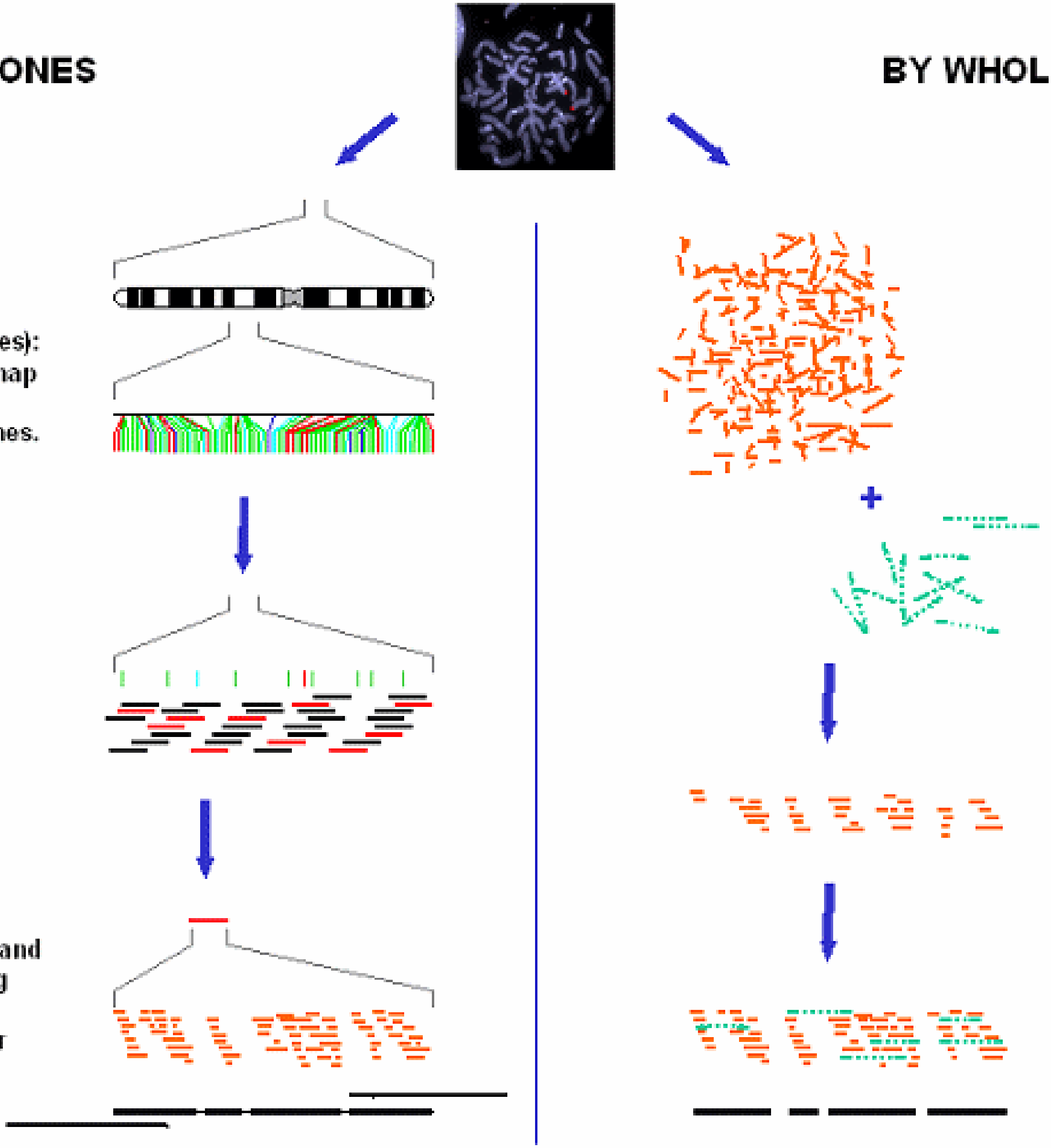
> Mouse genome and other primary model genomes

Differences in method and in starting material compared with human

Heterozygosity issues for assembly differ for inbred model organisms


HIERARCHICAL SHOTGUN
WHOLE-GENOME SHOTGUN
Genome
Random Reads
Assembly
Anchoring
Genome Assembly

STRATEGIES FOR SEQUENCING THE HUMAN GENOME



BY MAPPED CLONES

1. Construction of maps of ordered landmarks (genetic markers, genes): provides long-range map and organisation into individual chromosomes.

2. Physical maps of overlapping clones anchored to the landmark maps.

3. Selection of tile path (clones in red)

4. Shotgun sequencing and assembly (for working draft); subsequent directed finishing (for reference sequence).

BY WHOLE GENOME SHOTGUN

1. Shotgun sequencing of short-insert clones

2. Paired end sequencing of large-insert clones

3. Assembly of seed contigs (unitigs)

4. Incorporation of other sequences, and integration of long- range data.

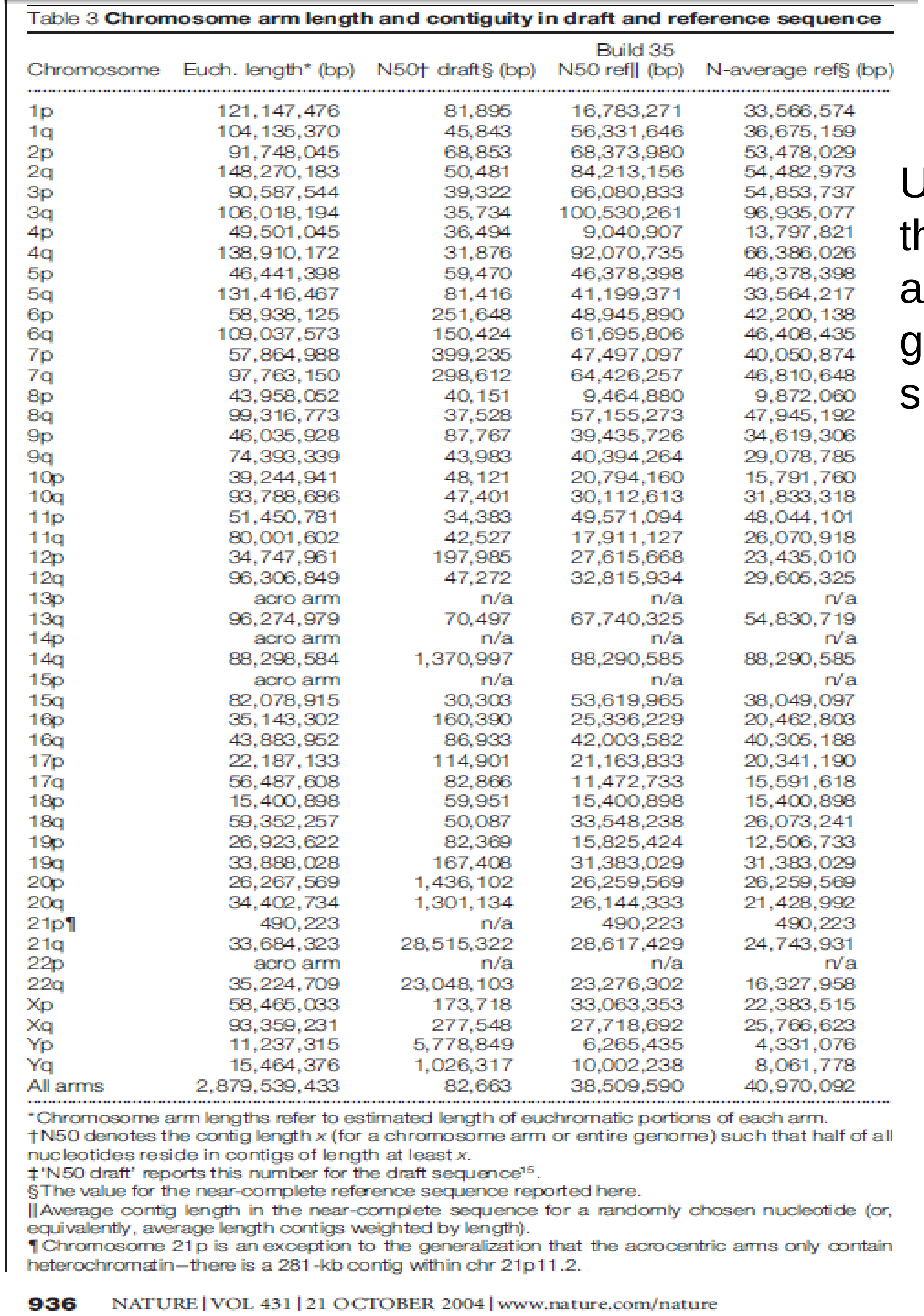


Table 3 **Chromosome arm length and contiguity in draft and reference sequence**

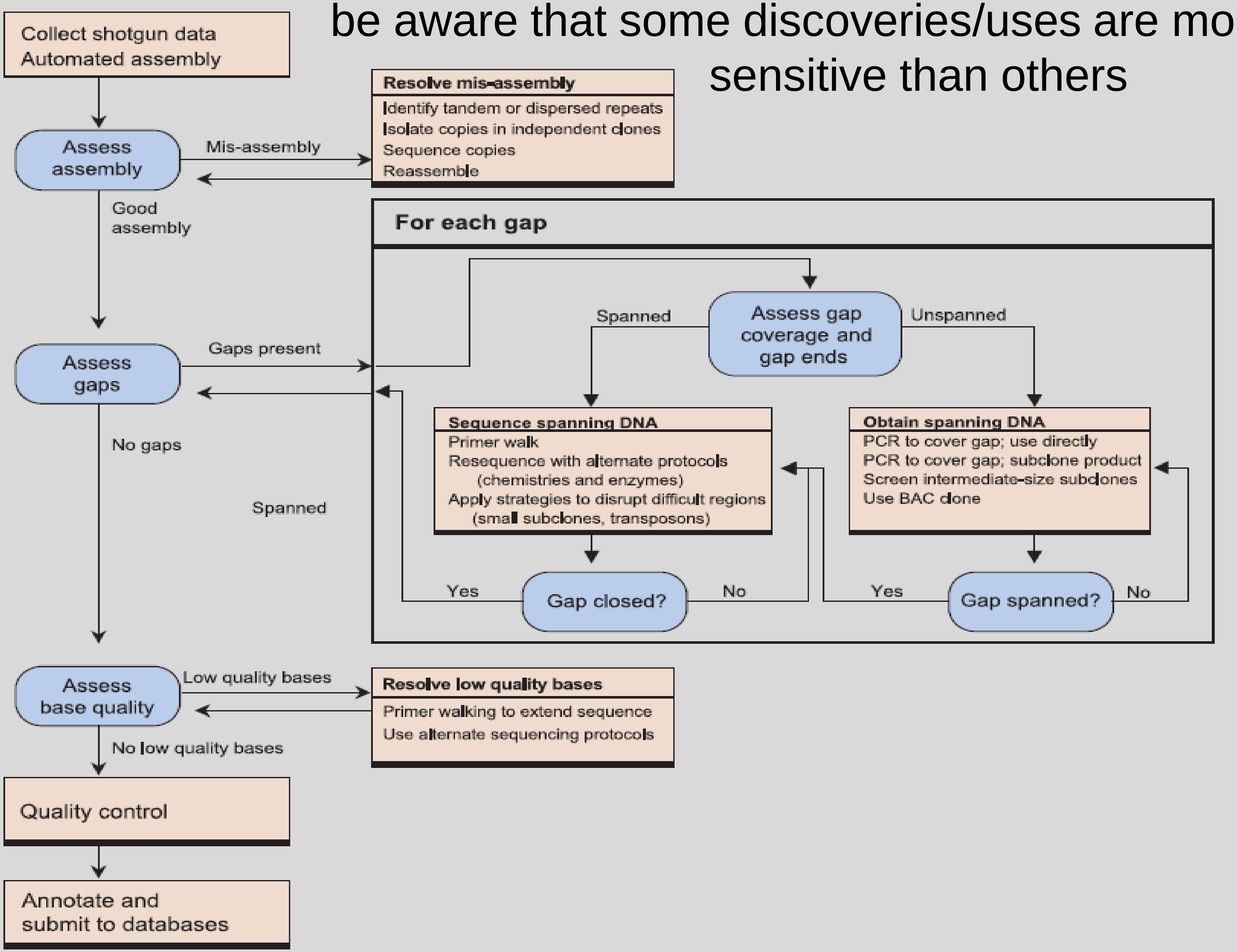
			Build 35	
Chromosome	Euch. length* (bp)	N50† draft§ (bp)	N50 ref|| (bp)	N-average ref§ (bp)
1p	121,147,476	81,895	16,783,271	33,566,574
1q	104,135,370	45,843	56,331,646	36,675,159
2p	91,748,045	68,853	68,373,980	53,478,029
2q	148,270,183	50,481	84,213,156	54,482,973
3p	90,587,544	39,322	66,080,833	54,853,737
3q	106,018,194	35,734	100,530,261	96,935,077
4p	49,501,045	36,494	9,040,907	13,797,821
4q	138,910,172	31,876	92,070,735	66,386,026
5p	46,441,398	59,470	46,378,398	46,378,398
5q	131,416,467	81,416	41,199,371	33,564,217
6p	58,938,125	251,648	48,945,890	42,200,138
6q	109,037,573	150,424	61,695,806	46,408,435
7p	57,864,988	399,235	47,497,097	40,050,874
7q	97,763,150	298,612	64,426,257	46,810,648
8p	43,958,052	40,151	9,464,880	9,872,060
8q	99,316,773	37,528	57,155,273	47,945,192
9p	46,035,928	87,767	39,435,726	34,619,306
9q	74,393,339	43,983	40,394,264	29,078,785
10p	39,244,941	48,121	20,794,160	15,791,760
10q	93,788,686	47,401	30,112,613	31,833,318
11p	51,450,781	34,383	49,571,094	48,044,101
11q	80,001,602	42,527	17,911,127	26,070,918
12p	34,747,961	197,985	27,615,668	23,435,010
12q	96,306,849	47,272	32,815,934	29,605,325
13p	acro arm	n/a	n/a	n/a
13q	96,274,979	70,497	67,740,325	54,830,719
14p	acro arm	n/a	n/a	n/a
14q	88,298,584	1,370,997	88,290,585	88,290,585
15p	acro arm	n/a	n/a	n/a
15q	82,078,915	30,303	53,619,965	38,049,097
16p	35,143,302	160,390	25,336,229	20,462,803
16q	43,883,952	86,933	42,003,582	40,305,188
17p	22,187,133	114,901	21,163,833	20,341,190
17q	56,487,608	82,866	11,472,733	15,591,618
18p	15,400,898	59,951	15,400,898	15,400,898
18q	59,352,257	50,087	33,548,238	26,073,241
19p	26,923,622	82,369	15,825,424	12,506,733
19q	33,888,028	167,408	31,383,029	31,383,029
20p	26,267,569	1,436,102	26,259,569	26,259,569
20q	34,402,734	1,301,134	26,144,333	21,428,992
21p¶	490,223	n/a	490,223	490,223
21q	33,684,323	28,515,322	28,617,429	24,743,931
22p	acro arm	n/a	n/a	n/a
22q	35,224,709	23,048,103	23,276,302	16,327,958
Xp	58,465,033	173,718	33,063,353	22,383,515
Xq	93,359,231	277,548	27,718,692	25,766,623
Yp	11,237,315	5,778,849	6,265,435	4,331,076
Yq	15,464,376	1,026,317	10,002,238	8,061,778
All arms	2,879,539,433	82,663	38,509,590	40,970,092

*Chromosome arm lengths refer to estimated length of euchromatic portions of each arm.
†N50 denotes the contig length *x* (for a chromosome arm or entire genome) such that half of all nucleotides reside in contigs of length at least *x*.
‡'N50 draft' reports this number for the draft sequence[15].
§The value for the near-complete reference sequence reported here.
||Average contig length in the near-complete sequence for a randomly chosen nucleotide (or, equivalently, average length contigs weighted by length).
¶Chromosome 21p is an exception to the generalization that the acrocentric arms only contain heterochromatin—there is a 281-kb contig within chr 21p11.2.

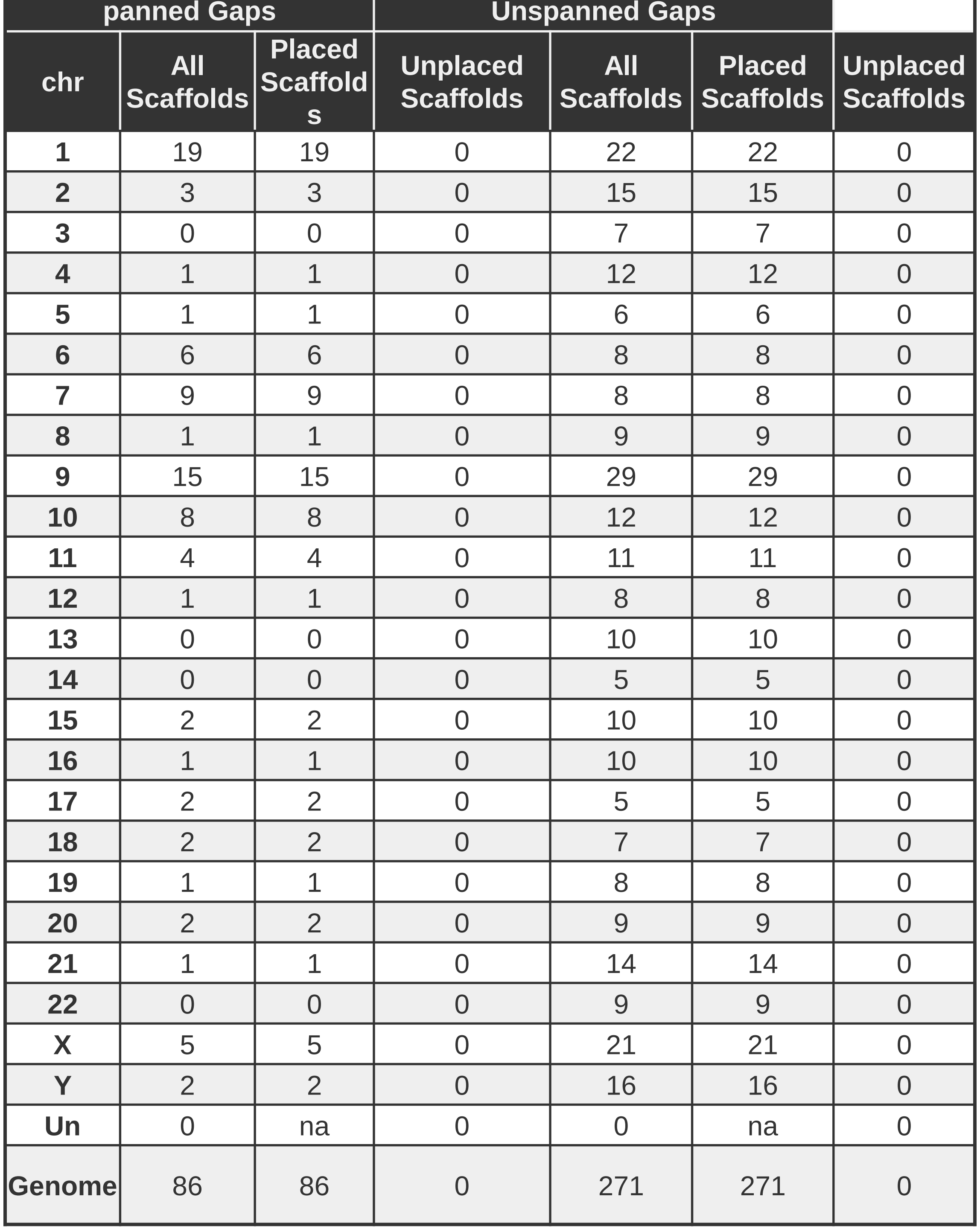
Useful metric: N50 =
the length in nucleotides
at which 50% of the assembled
genome is in blocks of the N50
size or longer

Details not important: Illustrating the additional HARD problem of achieving completeness and high quality in a genome sequence

be aware that some discoveries/uses are more quality sensitive than others


Collect shotgun data
Automated assembly
Assess assembly
Mis-assembly
Resolve mis-assembly
Identify tandem or dispersed repeats
Isolate copies in independent clones
Sequence copies
Reassemble
Good assembly
For each gap
Assess gaps
Gaps present
Spanned
Assess gap coverage and gap ends
Unspanned
Sequence spanning DNA
Primer walk
Resequence with alternate protocols (chemistries and enzymes)
Apply strategies to disrupt difficult regions (small subclones, transposons)
Obtain spanning DNA
PCR to cover gap; use directly
PCR to cover gap; subclone product
Screen intermediate-size subclones
Use BAC clone
No gaps
Spanned
Yes
Gap closed?
No
Yes
Gap spanned?
No
Assess base quality
Low quality bases
Resolve low quality bases
Primer walking to extend sequence
Use alternate sequencing protocols
No low quality bases
Quality control
Annotate and submit to databases

Box 2 Figure 1 Simplified flowchart for finishing of clones.



panned Gaps				Unspanned Gaps		
chr	All Scaffolds	Placed Scaffold s	Unplaced Scaffolds	All Scaffolds	Placed Scaffolds	Unplaced Scaffolds
1	19	19	0	22	22	0
2	3	3	0	15	15	0
3	0	0	0	7	7	0
4	1	1	0	12	12	0
5	1	1	0	6	6	0
6	6	6	0	8	8	0
7	9	9	0	8	8	0
8	1	1	0	9	9	0
9	15	15	0	29	29	0
10	8	8	0	12	12	0
11	4	4	0	11	11	0
12	1	1	0	8	8	0
13	0	0	0	10	10	0
14	0	0	0	5	5	0
15	2	2	0	10	10	0
16	1	1	0	10	10	0
17	2	2	0	5	5	0
18	2	2	0	7	7	0
19	1	1	0	8	8	0
20	2	2	0	9	9	0
21	1	1	0	14	14	0
22	0	0	0	9	9	0
X	5	5	0	21	21	0
Y	2	2	0	16	16	0
Un	0	na	0	0	na	0
Genome	86	86	0	271	271	0

Background information:

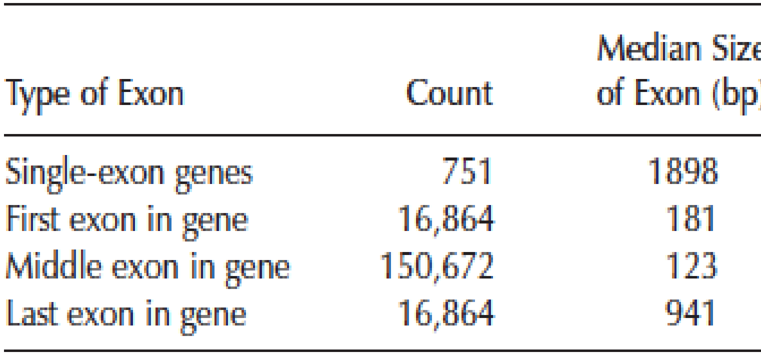
Distribution of GAPs in Current build of the human Genome

Gene types, functions and genome composition.
stats below for human are from one of several genome/ transcriptome
tations. Transcript isoform numbers and maps are complex and still not fully known.
Matters for debate about data and about importance

#BioType	Genes	Transcripts
IG_C_gene	16	18
IG_C_pseudogene	7	7
IG_D_gene	30	30
IG_J_gene	83	83
IG_J_pseudogene	3	3
IG_V_gene	180	181
IG_V_pseudogene	151	151
Mt_rRNA	2	2
Mt_tRNA	22	22
Mt_tRNA_pseudogene	580	580
TR_C_gene	3	3
TR_J_gene	13	13
TR_V_gene	48	48
TR_V_pseudogene	19	19
lincRNA	**1351**	**1592**
miRNA	**1756**	**1756**
miRNA_pseudogene	15	15
misc_RNA	1187	1187
misc_RNA_pseudogene	3	3
polymorphic_pseudogene	18	114
processed_transcript	9431	16068
protein_coding	**20540**	**118763**
pseudogene	10870	12595
rRNA	531	531
rRNA_pseudogene	179	179
scRNA_pseudogene	787	787
snRNA	1944	1944
snRNA_pseudogene	73	73
snoRNA	1521	1521
snoRNA_pseudogene	73	73
tRNA_pseudogene	128	128

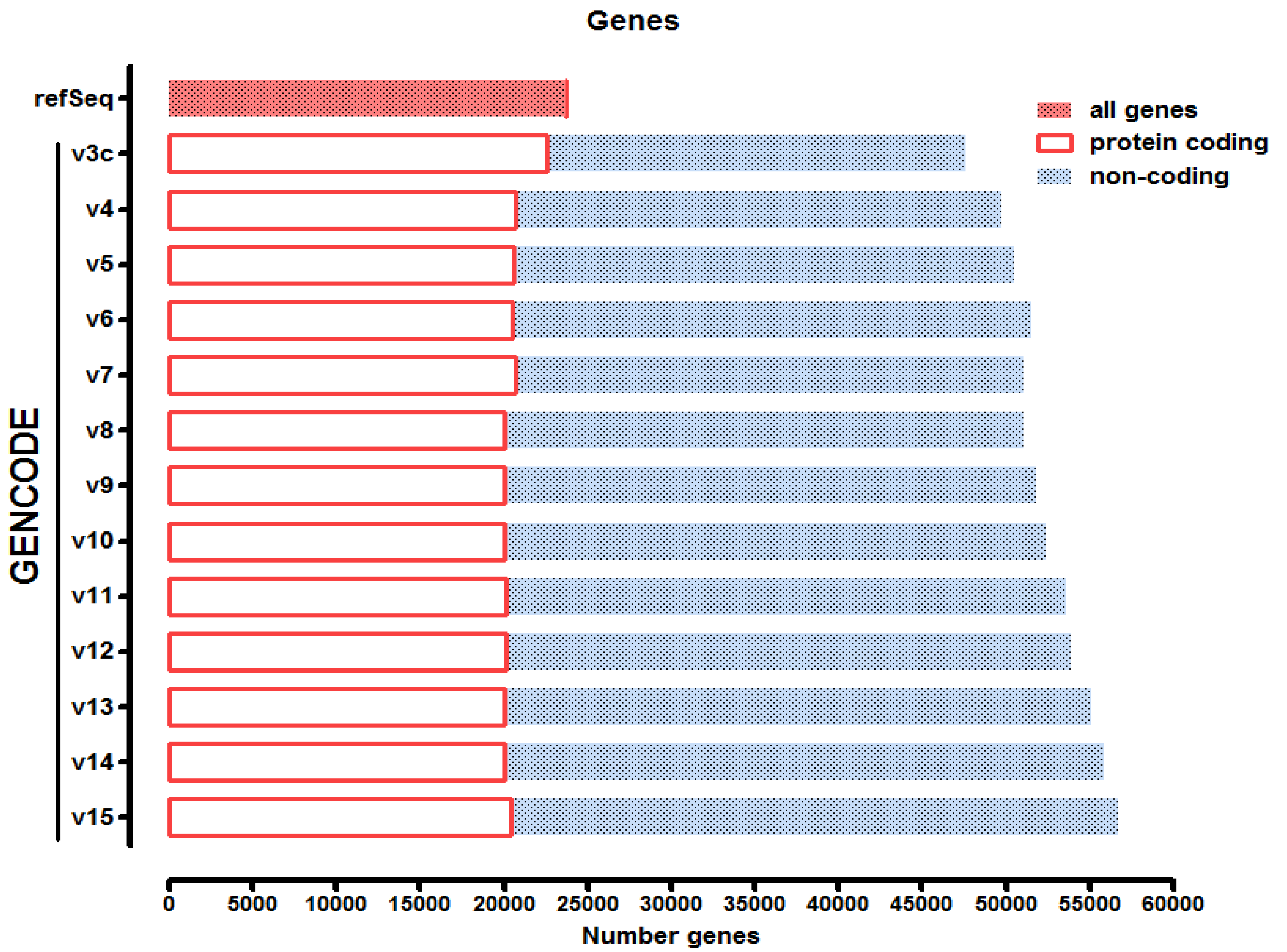
Anatomy of major gene class

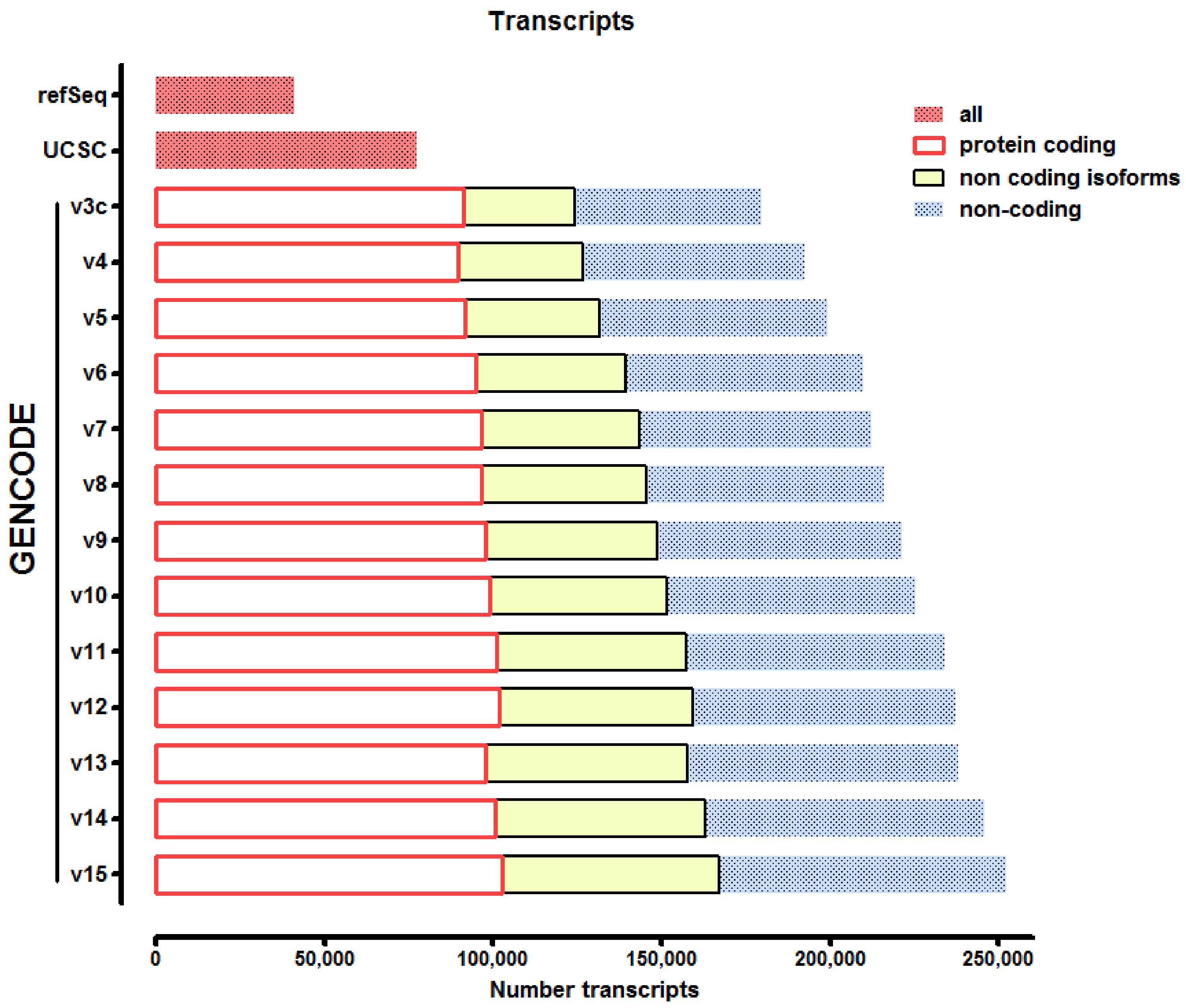
protein coding genes
median RNA coding length ~ 30Kb
median Exon number 8
median Exon lengths



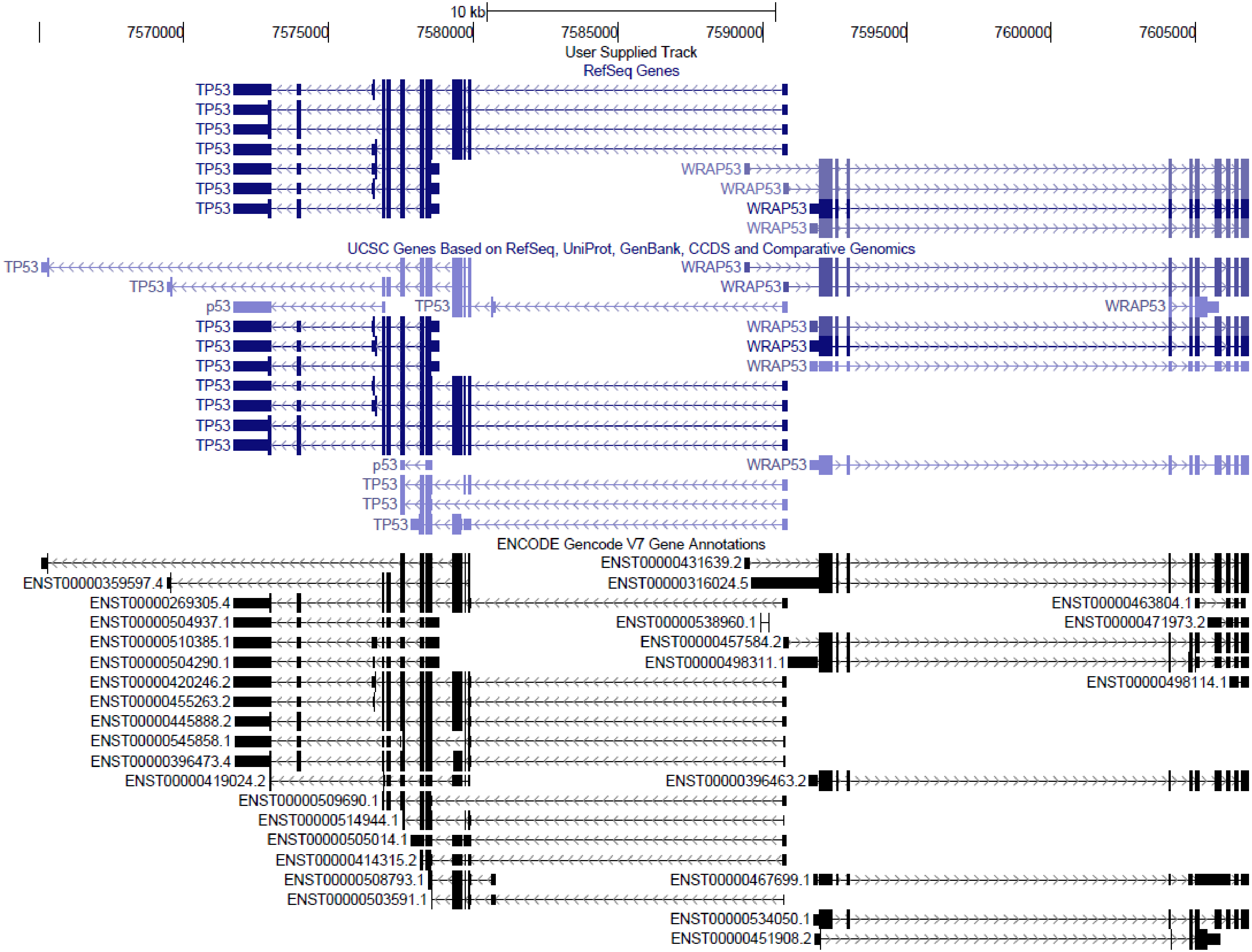
Type of Exon	Count	Median Size of Exon (bp
Single-exon genes	751	1898
First exon in gene	16,864	181
Middle exon in gene	150,672	123
Last exon in gene	16,864	941

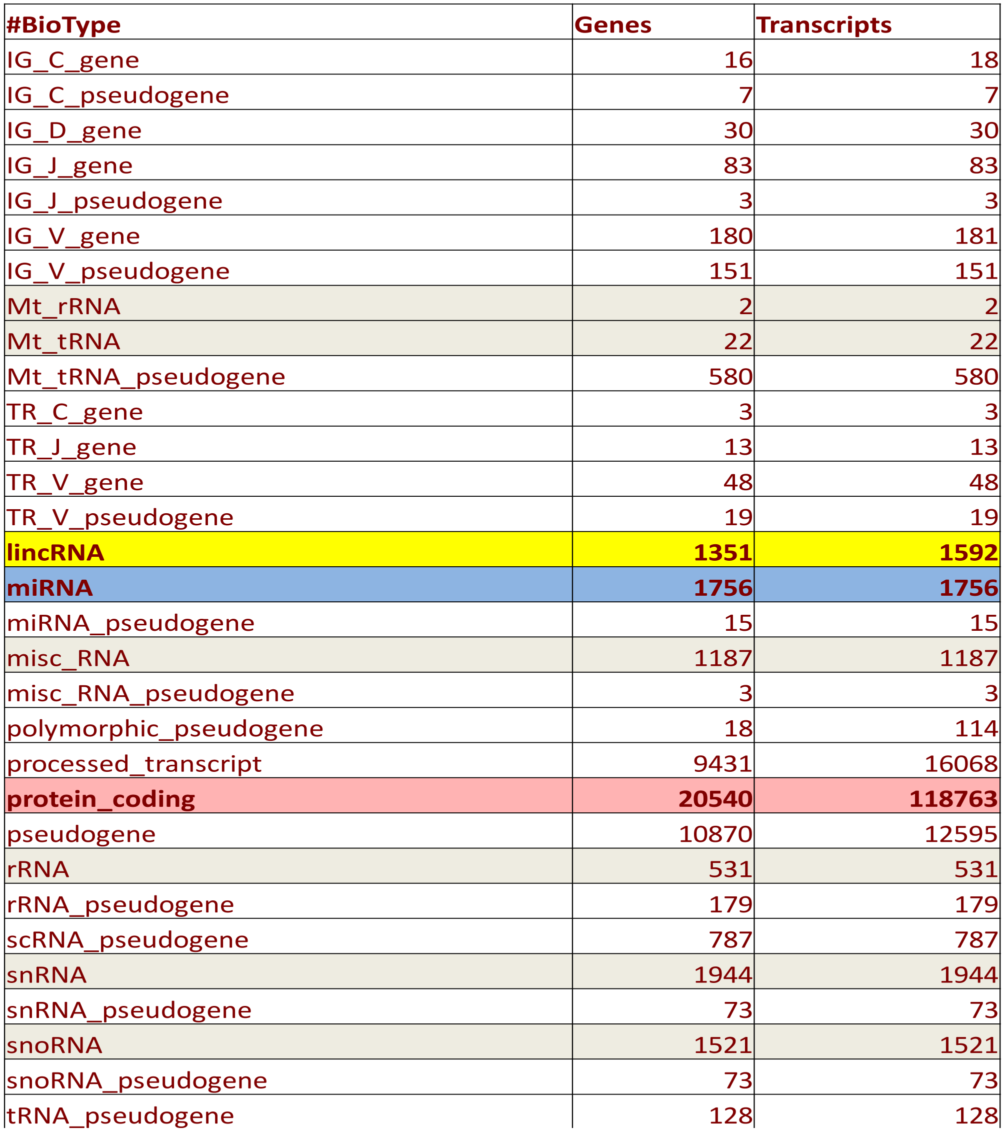
Different human genome “annotations” differ from each other and over time.
Biggest differences are in non-protein coding RNAs and their isoforms.




Transcripts
refSeq
UCSC
v3c
v4
v5
v6
v7
v8
v9
v10
v11
v12
v13
v14
v15
GENCODE
all
protein coding
non coding isoforms
non-coding
0
50,000
100,000
150,000
200,000
250,000
Number transcripts

“Typical gene” annotated per REFSEQ (top); UCSC middle; GENCODE V7 (bottom)





#BioType	Genes	Transcripts
IG_C_gene	16	18
IG_C_pseudogene	7	7
IG_D_gene	30	30
IG_J_gene	83	83
IG_J_pseudogene	3	3
IG_V_gene	180	181
IG_V_pseudogene	151	151
Mt_rRNA	2	2
Mt_tRNA	22	22
Mt_tRNA_pseudogene	580	580
TR_C_gene	3	3
TR_J_gene	13	13
TR_V_gene	48	48
TR_V_pseudogene	19	19
lincRNA	**1351**	**1592**
miRNA	**1756**	**1756**
miRNA_pseudogene	15	15
misc_RNA	1187	1187
misc_RNA_pseudogene	3	3
polymorphic_pseudogene	18	114
processed_transcript	9431	16068
protein_coding	**20540**	**118763**
pseudogene	10870	12595
rRNA	531	531
rRNA_pseudogene	179	179
scRNA_pseudogene	787	787
snRNA	1944	1944
snRNA_pseudogene	73	73
snoRNA	1521	1521
snoRNA_pseudogene	73	73
tRNA_pseudogene	128	128

Regarding pseudogenes:

Range of biological significance
Some expressed as RNA
Others not transcribed

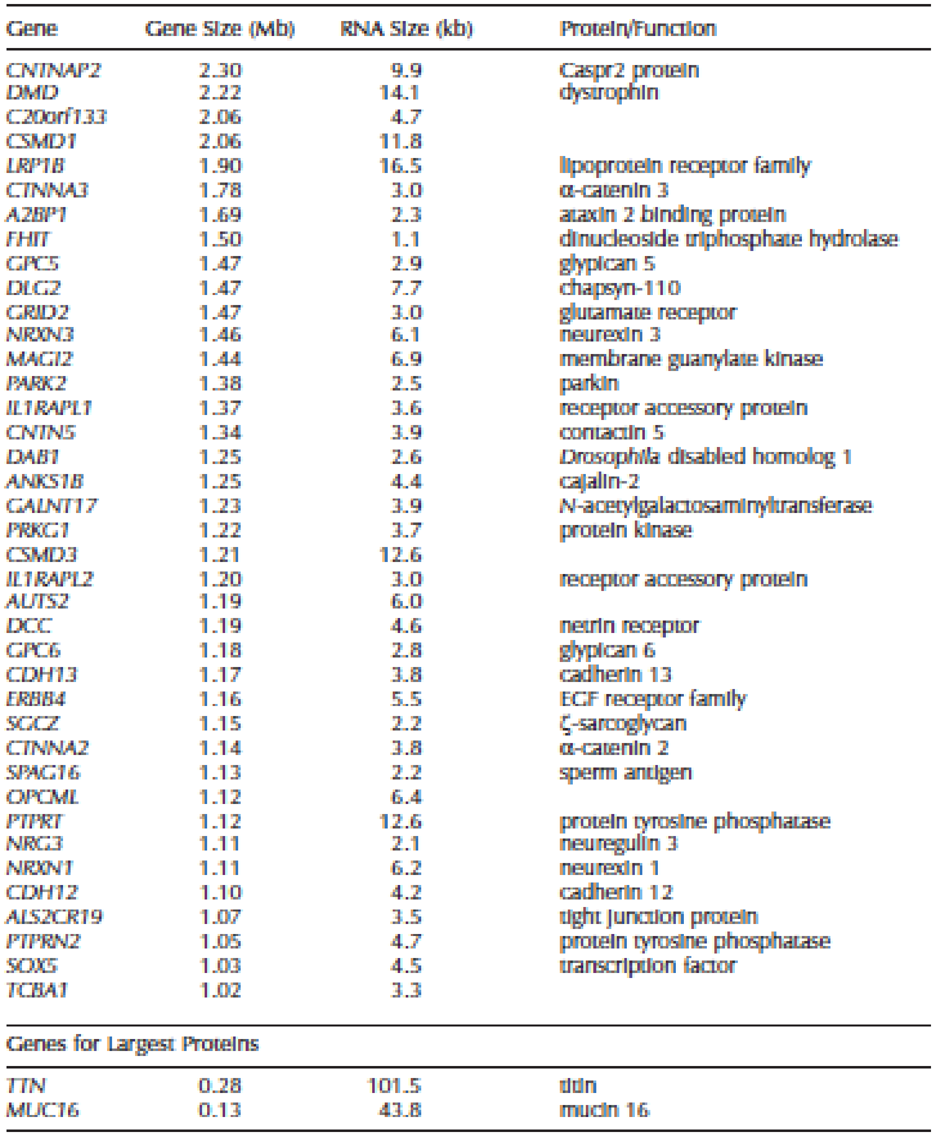
Major mechanisms of origin
1. duplication and mutation

2. “processed” retroposons
diagnostic = mRNA sequence

Implications of pseudogenes
for assays of gene expression
for assays of genomic sequence

>> never forget they are there
>> always ask if they are contributing to a genomic assay

Surveying the outliers: Big Genes



Gene	Gene Size (Mb)	RNA Size (kb)	Protein/Function
CNTNAP2	2.30	9.9	Caspr2 protein
DMD	2.22	14.1	dystrophin
C20orf133	2.06	4.7	
CSMD1	2.06	11.8	
LRP1B	1.90	16.5	lipoprotein receptor family
CTNNA3	1.78	3.0	α-catenin 3
A2BP1	1.69	2.3	ataxin 2 binding protein
FHIT	1.50	1.1	dinucleoside triphosphate hydrolase
GPC5	1.47	2.9	glypican 5
DLG2	1.47	7.7	chapsyn-110
GRID2	1.47	3.0	glutamate receptor
NRXN3	1.46	6.1	neurexin 3
MAGI2	1.44	6.9	membrane guanylate kinase
PARK2	1.38	2.5	parkin
IL1RAPL1	1.37	3.6	receptor accessory protein
CNTN5	1.34	3.9	contactin 5
DAB1	1.25	2.6	*Drosophila* disabled homolog 1
ANKS1B	1.25	4.4	cajalin-2
GALNT17	1.23	3.9	*N*-acetylgalactosaminyltransferase
PRKG1	1.22	3.7	protein kinase
CSMD3	1.21	12.6	
IL1RAPL2	1.20	3.0	receptor accessory protein
AUTS2	1.19	6.0	
DCC	1.19	4.6	netrin receptor
GPC6	1.18	2.8	glypican 6
CDH13	1.17	3.8	cadherin 13
ERBB4	1.16	5.5	EGF receptor family
SGCZ	1.15	2.2	ζ-sarcoglycan
CTNNA2	1.14	3.8	α-catenin 2
SPAG16	1.13	2.2	sperm antigen
OPCML	1.12	6.4	
PTPRT	1.12	12.6	protein tyrosine phosphatase
NRG3	1.11	2.1	neuregulin 3
NRXN1	1.11	6.2	neurexin 1
CDH12	1.10	4.2	cadherin 12
ALS2CR19	1.07	3.5	tight junction protein
PTPRN2	1.05	4.7	protein tyrosine phosphatase
SOX5	1.03	4.5	transcription factor
TCBA1	1.02	3.3	
Genes for Largest Proteins			
TTN	0.28	101.5	titin
MUC16	0.13	43.8	mucin 16

Implications for genetics:

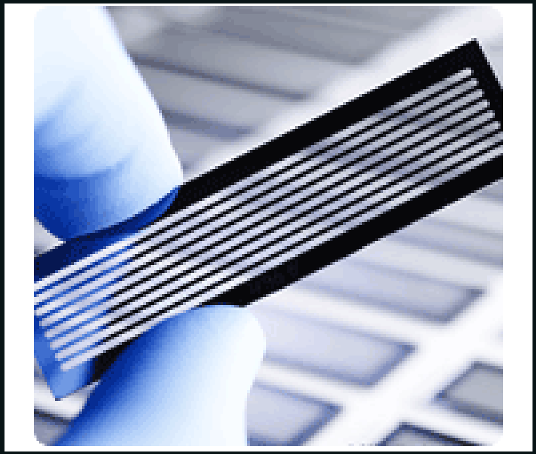
Big gene =
Big mutation
target.

Note dystrophin
A “pure” case example
because it is big;
recessive;
X-linked

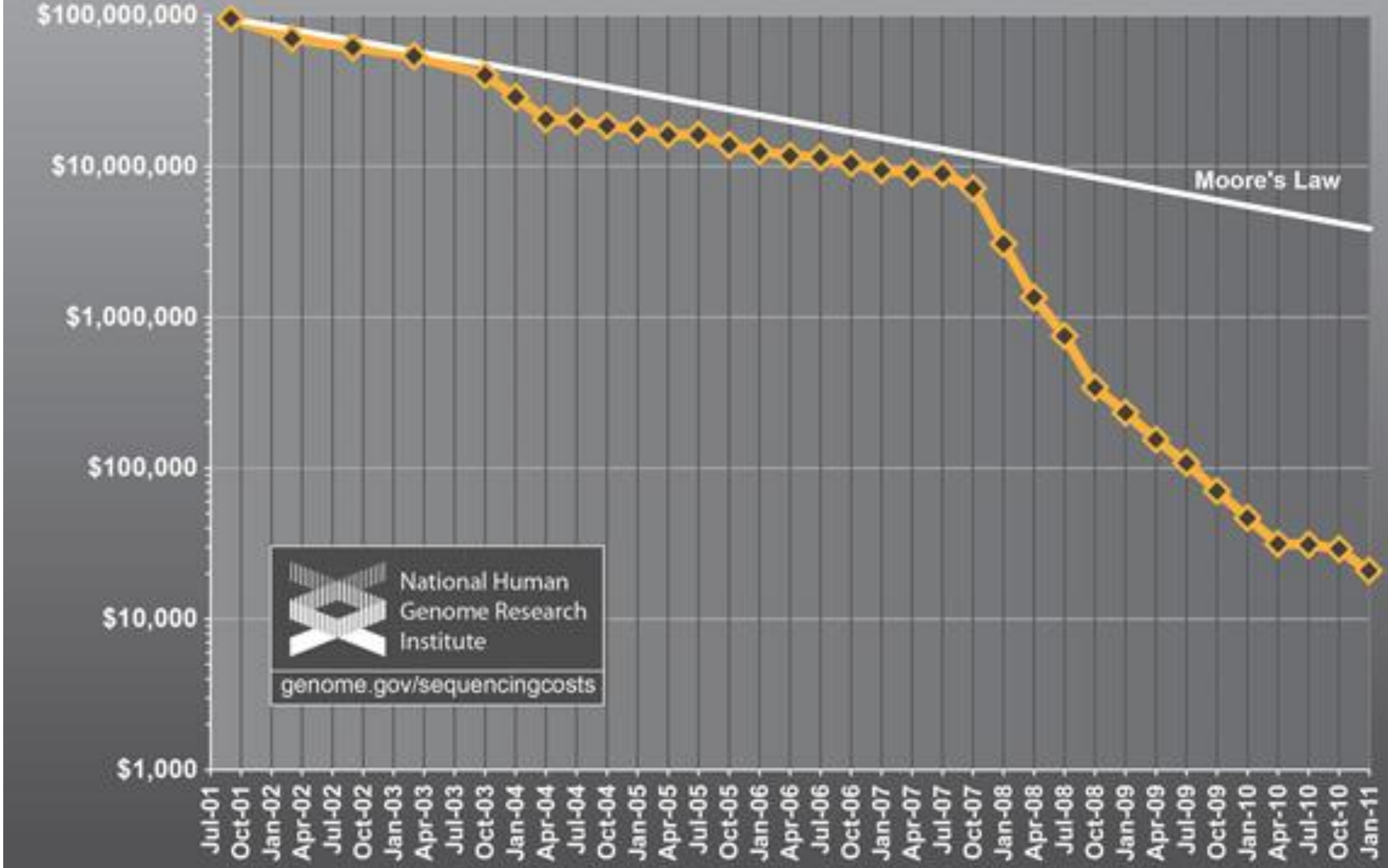
Technology has been rate-limiting: Basic DNA sequencing

1998 - Audacious goal for DNA sequencing

2 million bases/ year/ entire Project:	Accuracy ~10^{-4}	600 bp
2009 - 2- 4 billion bases/ 3 days/ machine:	Accuracy ~10^{-2}	25 bp
2011 - 200 billion / 6 days / machine:	Accuracy ~ 10^{-3}	2 x75 bp
2013 - 1-2 terabases / 3days / machine	Accuracy ~ 10^{-3}	2 x150 bp
bleeding edge Nanopore machines	*Accuracy?10^{-2}*	*>3,000 bp*



Cost per Genome



Which Increments of Technology matter for what problems?

1998 – Capillary electrophoresis machines (Hood, Smith, Hunkapillar CIT/ABI)
2 million bases/ year/ entire Project: Accuracy ~10^{-4} 600 bp
*Made plausible the previously unrealistic goal for human genome

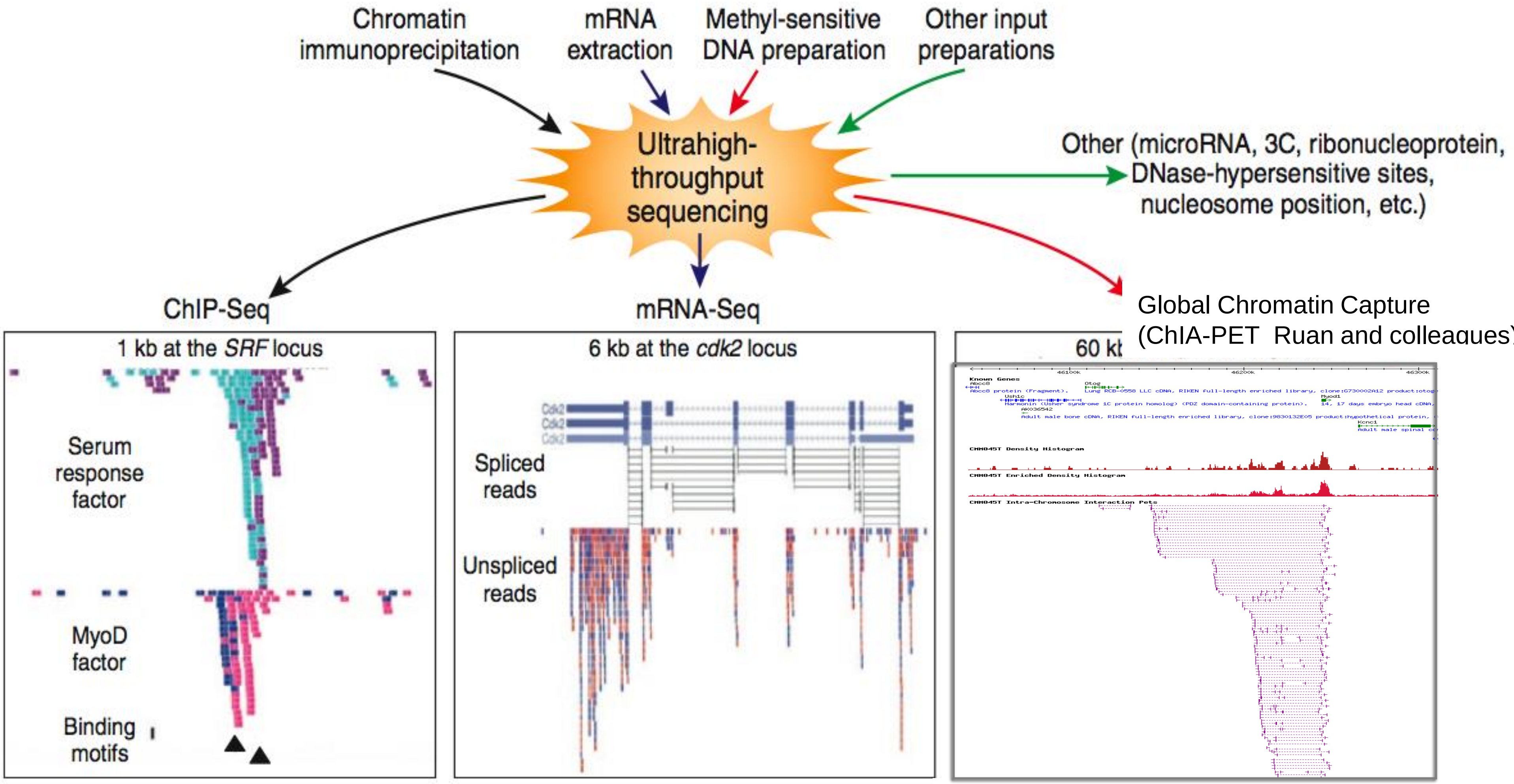
2007 - 2-4 billion bases/ 3 days/ machine: Accuracy ~10^{-2} 25 bp
First Solexa/Illumina “short-read” machines
*Made *comprehensive genome-wide* assays possible for big genomes
Previously limited mainly to yeast (summarized in Wold and Myers, 2008)

2011 - 200 billion / 6 days / machine: Accuracy ~ 10^{-3} 2 x75 bp
Made RNA isoforms plausible (still imperfect)

2013 - 1-2 terabases / 3days / machine Accuracy ~ 10^{-3} 2 x150 bp
*Made possible clinical sequencing turnaround ~$5,000 per patient

bleeding edge Nanopore machines Accuracy?10^{-2} >3,000 bp
You predict the impact.......and prepare to discuss

DNA sequencing became routine method of quantitative assays for many experiment types where RNA or DNA is the substrate, and especially where sub-portions of genome are enriched


Chromatin immunoprecipitation
mRNA extraction
Methyl-sensitive DNA preparation
Other input preparations
Ultrahigh-throughput sequencing
Other (microRNA, 3C, ribonucleoprotein, DNase-hypersensitive sites, nucleosome position, etc.)
ChIP-Seq
mRNA-Seq
1 kb at the SRF locus
Serum response factor
MyoD factor
Binding motifs
6 kb at the cdk2 locus
Cdk2
Cdk2
Cdk2
Spliced reads
Unspliced reads
60 kb
46100k
46200k
46300k
Known Genes
Abcc8
Abcc8 protein (Fragment).
Otog
Lung RCB-0558 LLC cDNA, RIKEN full-length enriched library, clone:G730002A12 product:otog
Ush1c
Harmonin (Usher syndrome 1C protein homolog) (PDZ domain-containing protein).
Myod1
14, 17 days embryo head cDNA,
AK036542
Adult male bone cDNA, RIKEN full-length enriched library, clone:9830132E05 product:hypothetical protein,
Kcnc1
Adult male spinal co
CHM045T Density Histogram
CHM045T Enriched Density Histogram
CHM045T Intra-Chromosome Interaction Pets

Global Chromatin Capture
(ChIA-PET Ruan and colleagues)

Factors, polymerase load
chromatin marks
DNA motif discovery

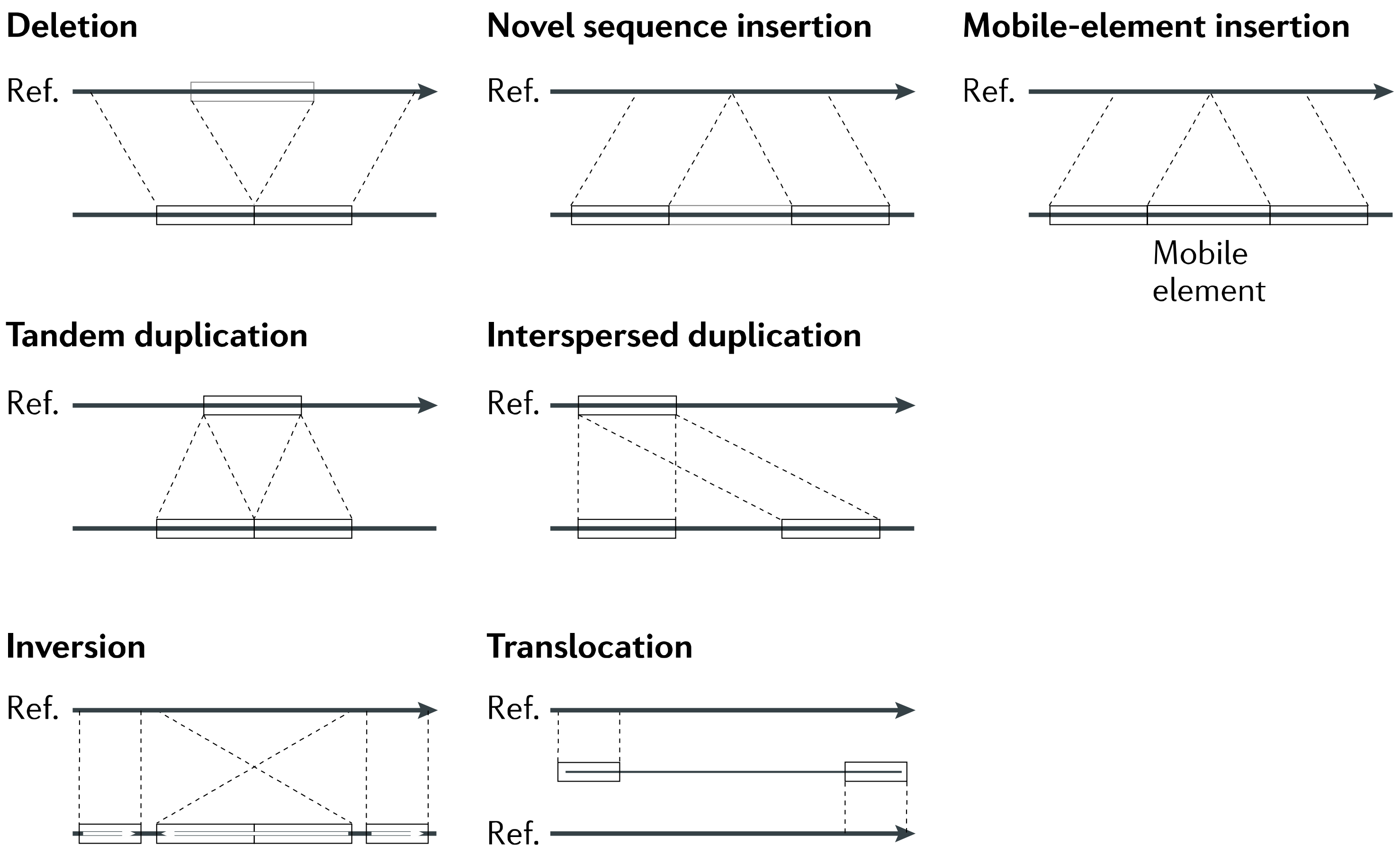
Genes
expressed
Isoforms defined

Long distance connecti

Who consorts
with whom?

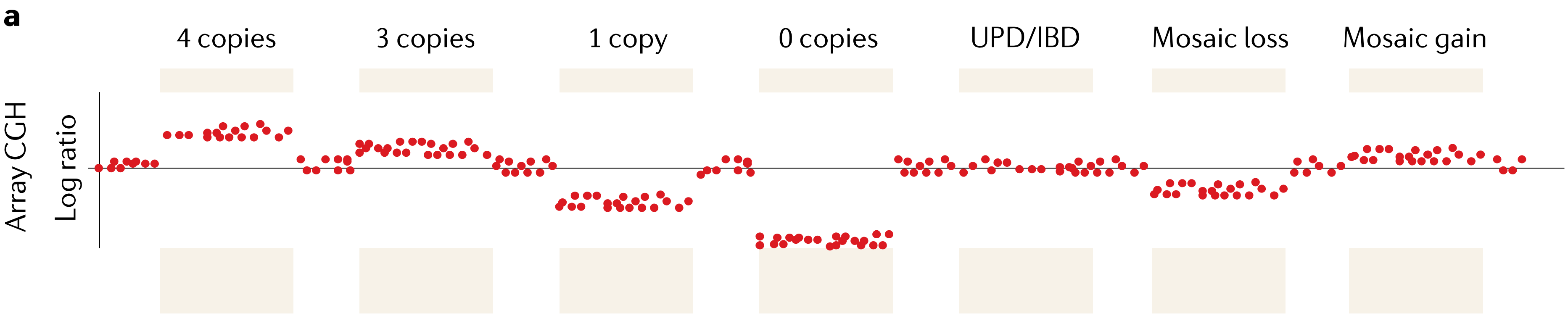
CHM045T Cluster Inter Interactions

Human genome variation - *Much* more than SNPs
Structural Variation is the general terminology



Copy number variation is a common and important consequence
CNV = differences in number for a gene or other sequence

How is CNV detected experimentally? Multiple ways by now –
differing issues of sensitivity, noise, resolution



Evan Eichler and colleagues; data via microarray CGH
Array hybridization convention is log 2 ratio probe a/b]

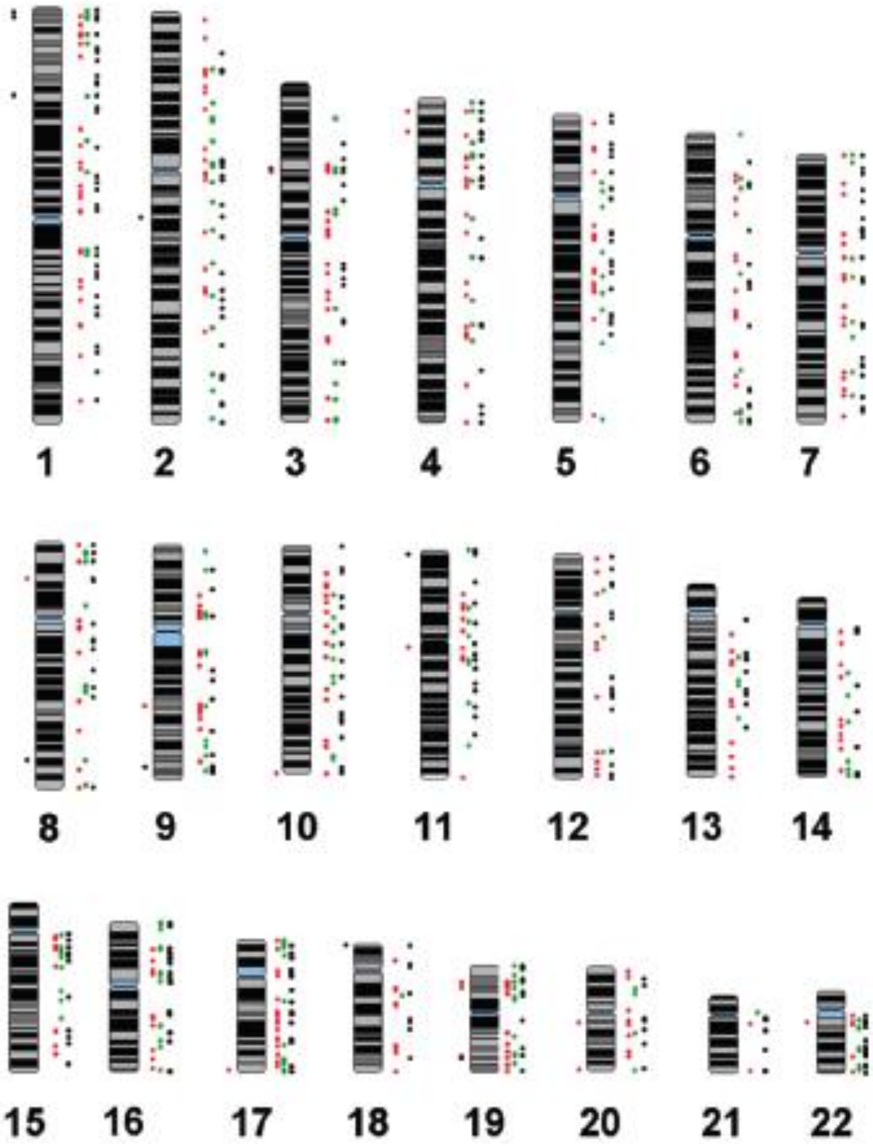
Human Segmental Duplication Map

implications – functional and technical - for individual genomics

- 1kb to 500kb size
>90% similar

- 2 - 6 copies (up to 20)

>5% of genome



igure 6. Distribution of CNV clones. High-frequency CNV clones are shown as dots to the right of each chromosome; red, green, and lack dots represent presence in three, four or five, and six or more individuals, respectively. Dots to the left of the chromosomes epresent locations of CNVs that overlap microRNAs (*red dots*) and select cancer genes (*black dots*).

Overall map shows better the range of sizes; the telomeric and centromeric biases

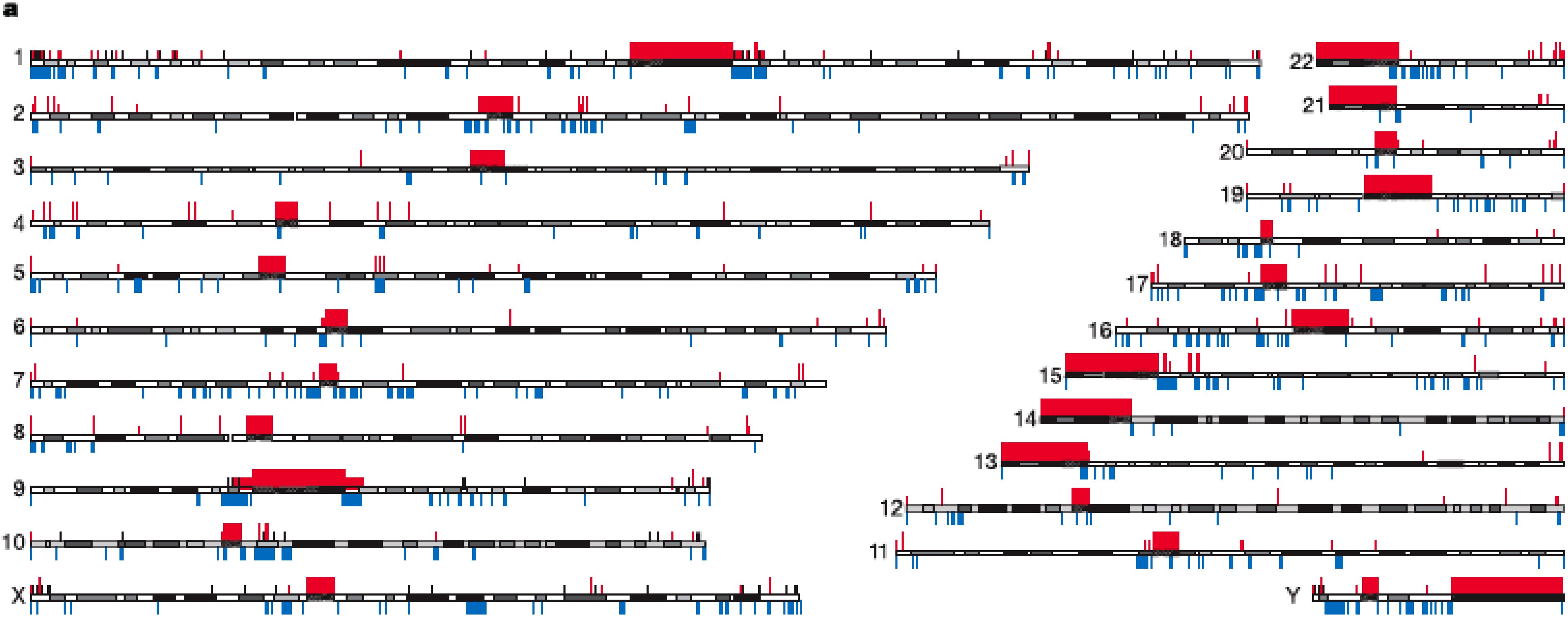
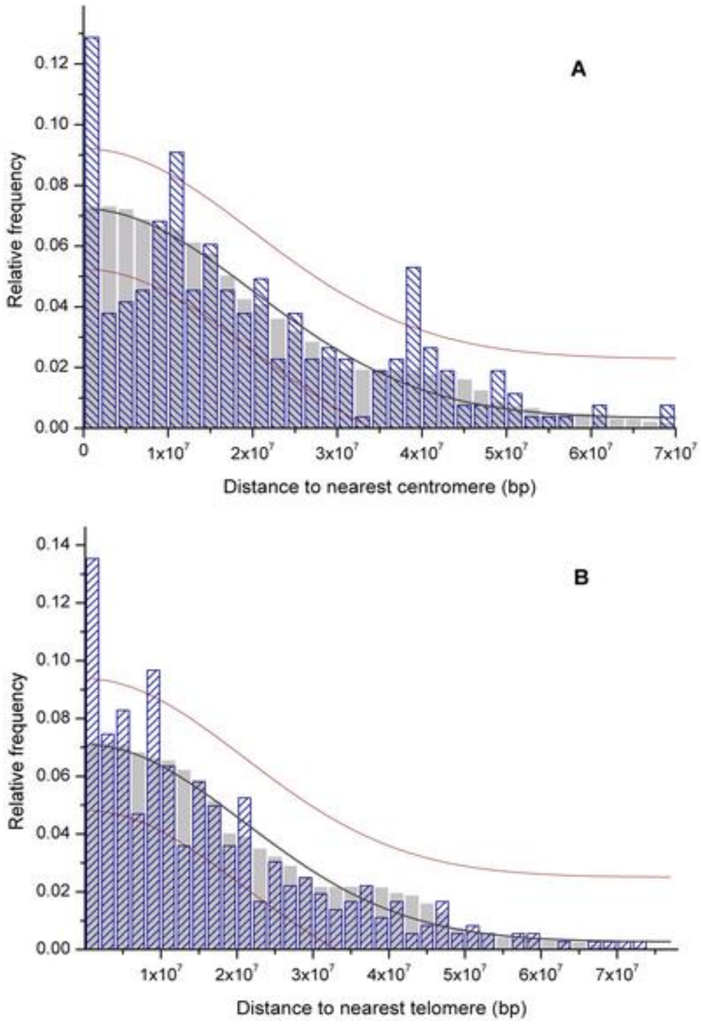


Figure 1. Relative Frequency Histograms of Distances from Human CNVs to the Nearest Centromere or Telomere



Nguyen D-Q, Webber C, Ponting CP (2006) Bias of Selection on Human Copy-Number Variants. PLoS Genet 2(2): e20. doi:10.1371/journal.pgen.0020020

http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.0020020

Specific Example: Pleiotropic Skeletal Malformations due to duplications of part of Indian HedgeHog (IHH)

- Example of dominant duplication disorder.   Because of previously unappreciated function of *ihh* in signaling in bone development, this explains heritable malformation at multiple body sites

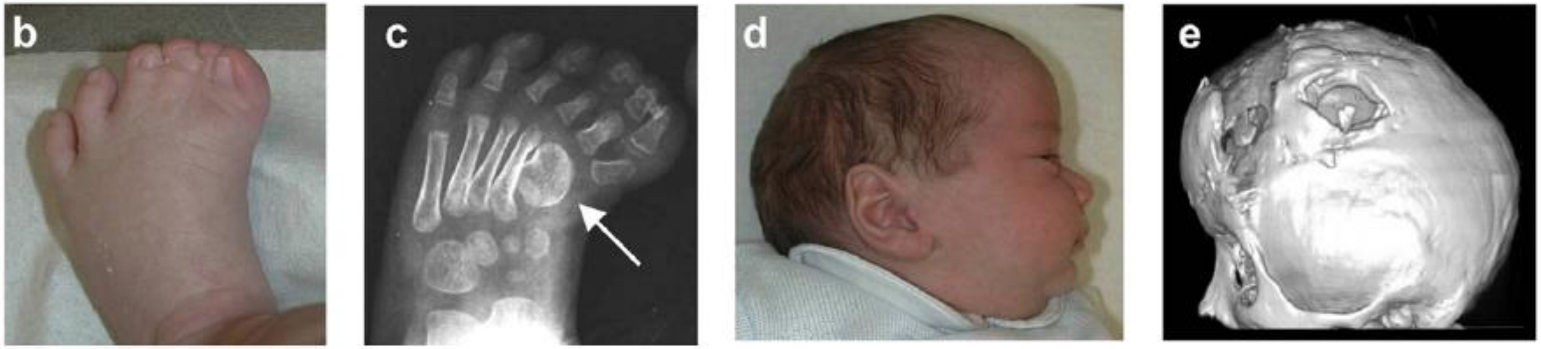
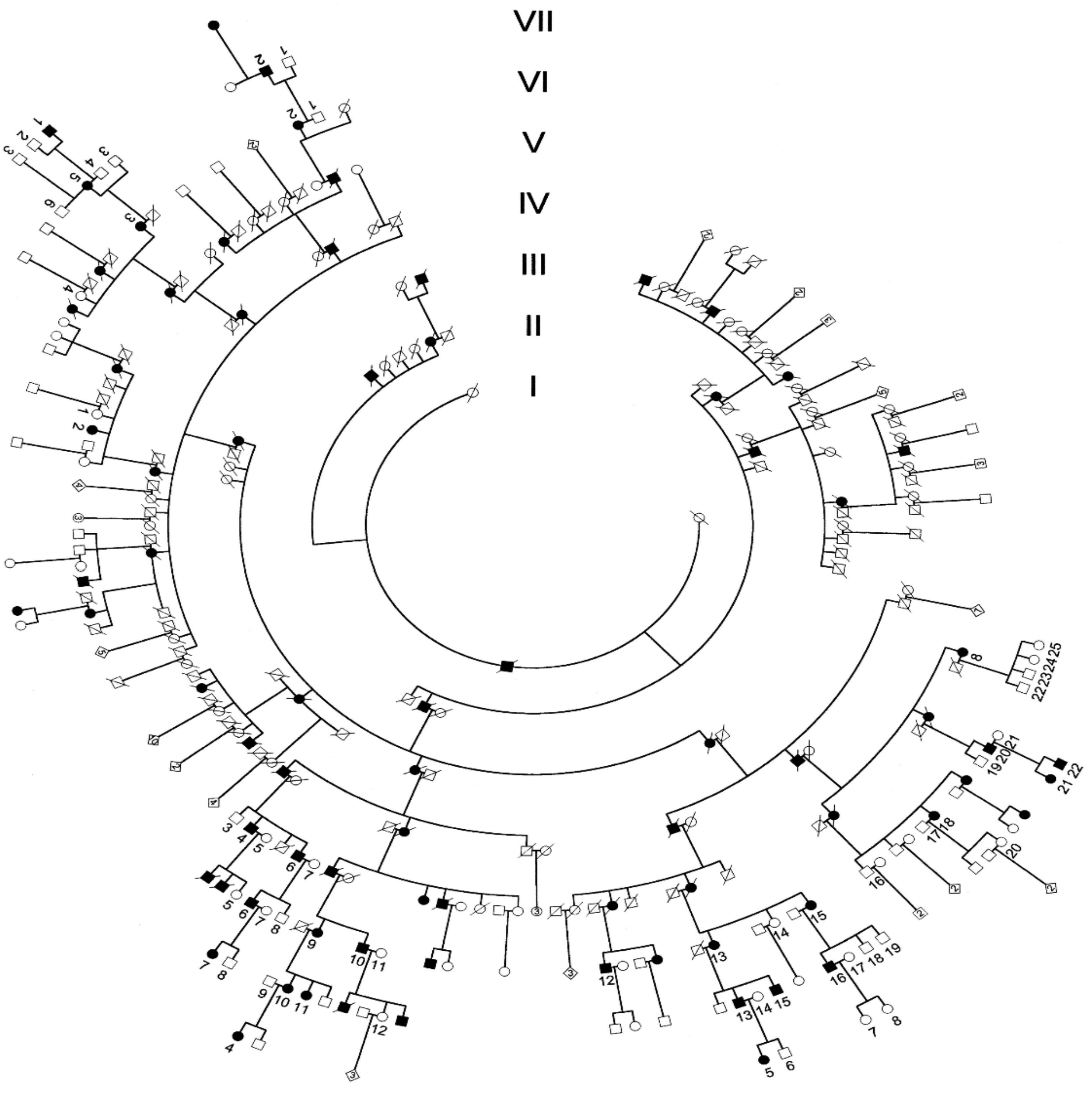


Figure 1.   CNVs at the IHH Locus on 2q35 and the Associated Clinical Phenotype

Klopacki et al, 2011 Am J. Hum Genetics

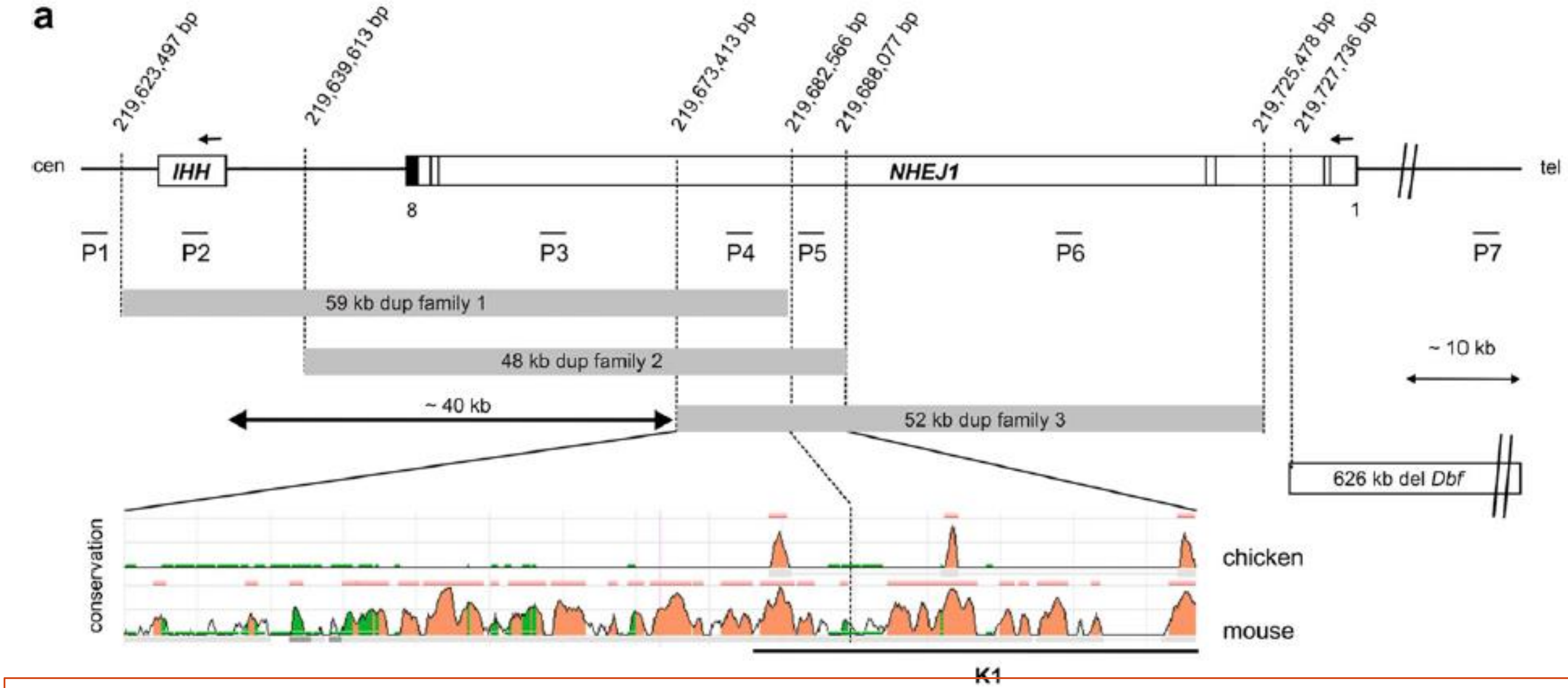
- Ilustrates the impact of duplication of distant cis-acting regulatory component(s) of a gene (buried within a second unrelated gene (*nhej*). [Interpretation issues on account of this]
- Next: How do you move from a mapped human locus to build and test a hypothesis of regulatory element causation?



Bosse et al., 2000 *Am. J.*

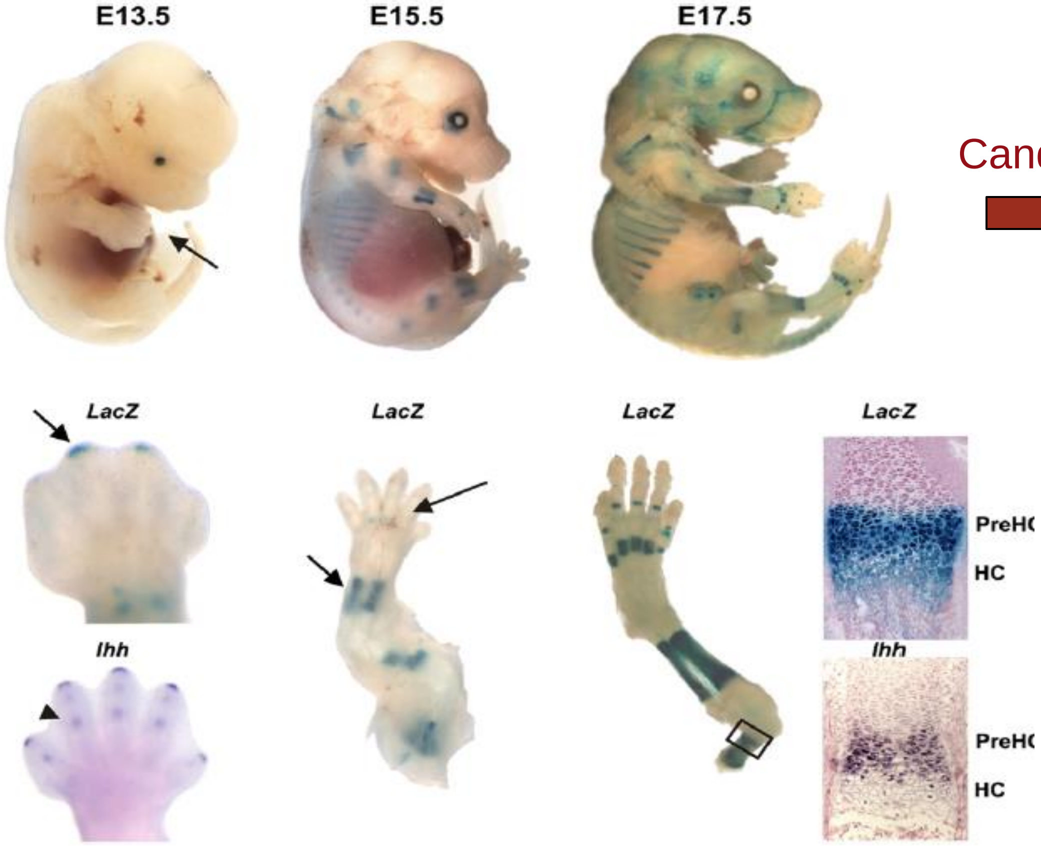
ihh duplication structures – 3 distinct families

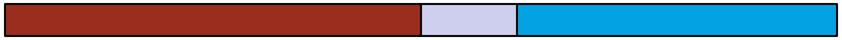
Duplications at distant cis-acting regulatory component of a gene (buried within a second unrelated gene (*nhej*). Interpretation issues on account of this. Note highly conserved noncoding sequence within region P4 (and in segment K1 of mouse).



How would you test the hypothesis that it is altered expression of ihh, attributable to CRMs (cis-regulatory modules composed of transcriptional enhancers/silencers) that is causal? What piece of DNA would you test, based on the above map? Why?

Indian Hedgehog [paralog of Sonic Hedgehog (shh)] regulatory sequence: test for domains of action in mouse

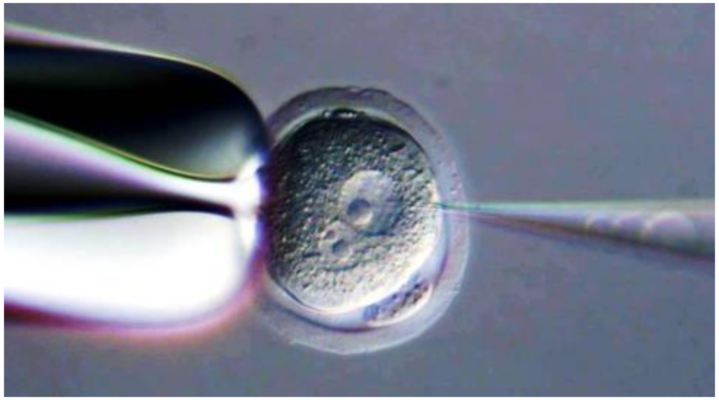




Candidate reg seq K1

Reporter (lacz)

basal
promoter



Candidate biological significance groups for CNV consider the group of tumor suppressor genes and oncogenes

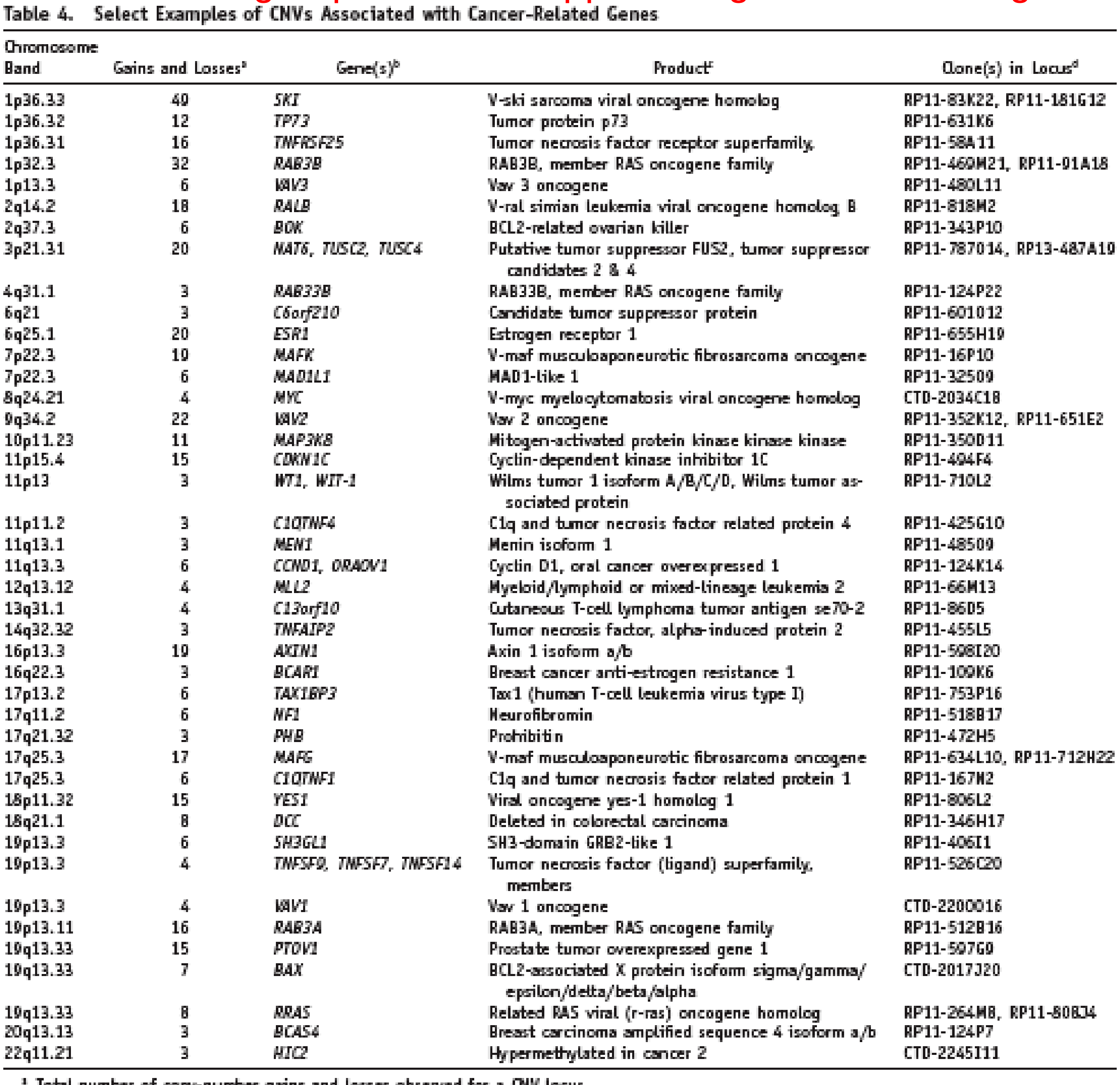


Table 4. Select Examples of CNVs Associated with Cancer-Related Genes

Chromosome Band	Gains and Losses[a]	Gene(s)[b]	Product[c]	Clone(s) in Locus[d]
1p36.33	49	*SKI*	V-ski sarcoma viral oncogene homolog	RP11-83K22, RP11-181G12
1p36.32	12	*TP73*	Tumor protein p73	RP11-631K6
1p36.31	16	*TNFRSF25*	Tumor necrosis factor receptor superfamily,	RP11-58A11
1p32.3	32	*RAB3B*	RAB3B, member RAS oncogene family	RP11-469M21, RP11-91A18
1p13.3	6	*VAV3*	Vav 3 oncogene	RP11-480L11
2q14.2	18	*RALB*	V-ral simian leukemia viral oncogene homolog B	RP11-818M2
2q37.3	6	*BOK*	BCL2-related ovarian killer	RP11-343P10
3p21.31	20	*NAT6, TUSC2, TUSC4*	Putative tumor suppressor FUS2, tumor suppressor candidates 2 & 4	RP11-787O14, RP13-487A19
4q31.1	3	*RAB33B*	RAB33B, member RAS oncogene family	RP11-124P22
6q21	3	*C6orf210*	Candidate tumor suppressor protein	RP11-601O12
6q25.1	20	*ESR1*	Estrogen receptor 1	RP11-655H19
7p22.3	19	*MAFK*	V-maf musculoaponeurotic fibrosarcoma oncogene	RP11-16P10
7p22.3	6	*MAD1L1*	MAD1-like 1	RP11-32509
8q24.21	4	*MYC*	V-myc myelocytomatosis viral oncogene homolog	CTD-2034C18
9q34.2	22	*VAV2*	Vav 2 oncogene	RP11-352K12, RP11-651E2
10p11.23	11	*MAP3K8*	Mitogen-activated protein kinase kinase kinase	RP11-350O11
11p15.4	15	*CDKN1C*	Cyclin-dependent kinase inhibitor 1C	RP11-494F4
11p13	3	*WT1, WIT-1*	Wilms tumor 1 isoform A/B/C/D, Wilms tumor associated protein	RP11-710L2
11p11.2	3	*C1QTNF4*	C1q and tumor necrosis factor related protein 4	RP11-425G10
11q13.1	3	*MEN1*	Menin isoform 1	RP11-48509
11q13.3	6	*CCND1, ORAOV1*	Cyclin D1, oral cancer overexpressed 1	RP11-124K14
12q13.12	4	*MLL2*	Myeloid/lymphoid or mixed-lineage leukemia 2	RP11-66M13
13q31.1	4	*C13orf10*	Cutaneous T-cell lymphoma tumor antigen *se70-2*	RP11-86O5
14q32.32	3	*TNFAIP2*	Tumor necrosis factor, alpha-induced protein 2	RP11-455L5
16p13.3	19	*AXIN1*	Axin 1 isoform a/b	RP11-598I20
16q22.3	3	*BCAR1*	Breast cancer anti-estrogen resistance 1	RP11-109K6
17p13.2	6	*TAX1BP3*	Tax1 (human T-cell leukemia virus type I)	RP11-753P16
17q11.2	6	*NF1*	Neurofibromin	RP11-518B17
17q21.32	3	*PHB*	Prohibitin	RP11-472H5
17q25.3	17	*MAFG*	V-maf musculoaponeurotic fibrosarcoma oncogene	RP11-634L10, RP11-712H22
17q25.3	6	*C1QTNF1*	C1q and tumor necrosis factor related protein 1	RP11-167N2
18p11.32	15	*YES1*	Viral oncogene yes-1 homolog 1	RP11-806L2
18q21.1	8	*DCC*	Deleted in colorectal carcinoma	RP11-346H17
19p13.3	6	*SH3GL1*	SH3-domain GRB2-like 1	RP11-406I1
19p13.3	4	*TNFSF9, TNFSF7, TNFSF14*	Tumor necrosis factor (ligand) superfamily, members	RP11-526C20
19p13.3	4	*VAV1*	Vav 1 oncogene	CTD-2200O16
19p13.11	16	*RAB3A*	RAB3A, member RAS oncogene family	RP11-512B16
19q13.33	15	*PTOV1*	Prostate tumor overexpressed gene 1	RP11-597G9
19q13.33	7	*BAX*	BCL2-associated X protein isoform sigma/gamma/epsilon/delta/beta/alpha	CTD-2017J20
19q13.33	8	*RRAS*	Related RAS viral (r-ras) oncogene homolog	RP11-264M8, RP11-808J4
20q13.13	3	*BCAS4*	Breast carcinoma amplified sequence 4 isoform a/b	RP11-124P7
22q11.21	3	*HIC2*	Hypermethylated in cancer 2	CTD-2245I11

[a] Total number of copy-number gains and losses observed for a CNV locus.

Sensory genes – early list – concept is the point

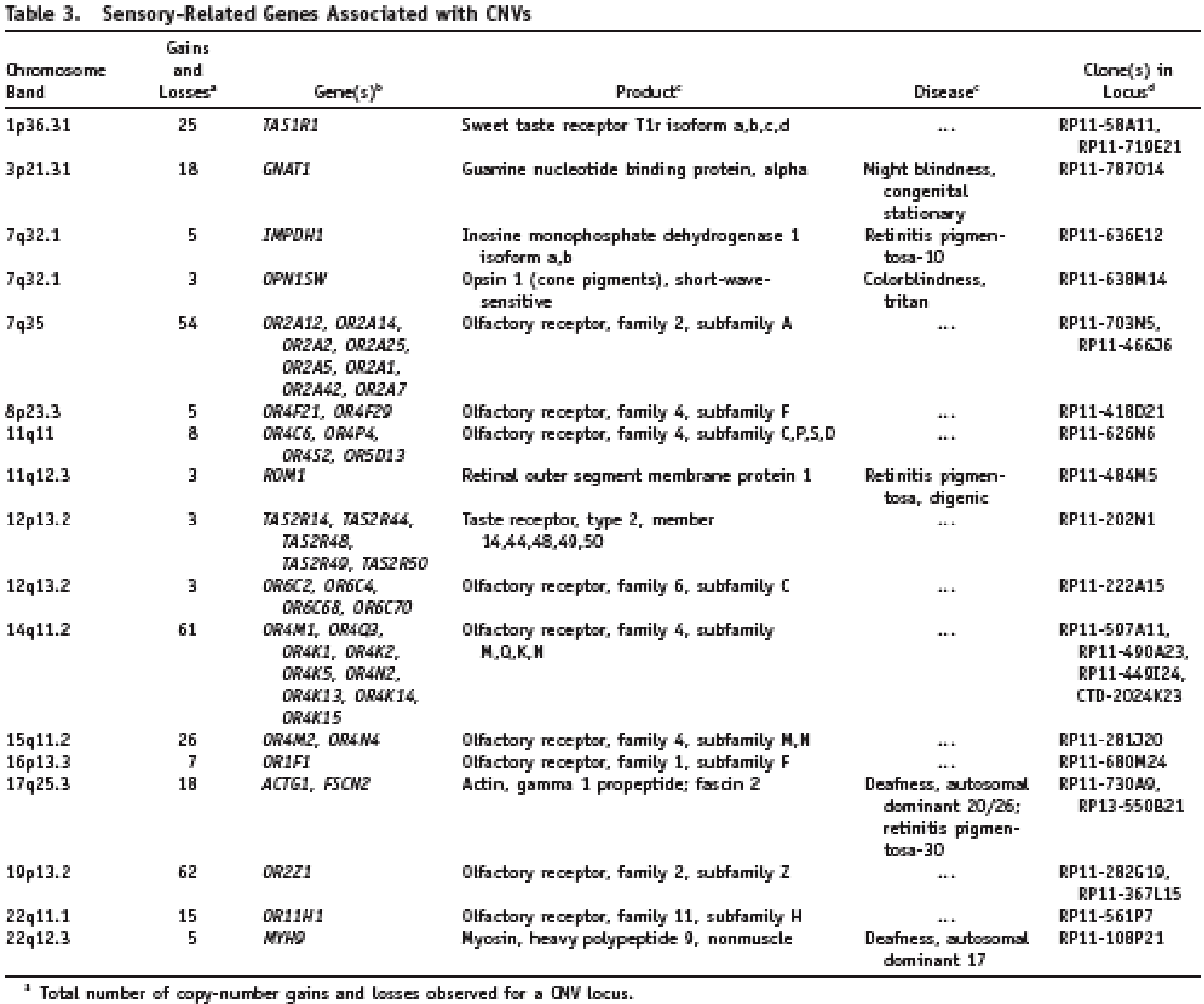


Table 3. Sensory-Related Genes Associated with CNVs

Chromosome Band	Gains and Losses[a]	Gene(s)[b]	Product[c]	Disease[c]	Clone(s) in Locus[d]
1p36.31	25	*TAS1R1*	Sweet taste receptor T1r isoform a,b,c,d	...	RP11-58A11, RP11-719E21
3p21.31	18	*GNAT1*	Guanine nucleotide binding protein, alpha	Night blindness, congenital stationary	RP11-787O14
7q32.1	5	*IMPDH1*	Inosine monophosphate dehydrogenase 1 isoform a,b	Retinitis pigmentosa-10	RP11-636E12
7q32.1	3	*OPN1SW*	Opsin 1 (cone pigments), short-wave-sensitive	Colorblindness, tritan	RP11-638M14
7q35	54	*OR2A12, OR2A14, OR2A2, OR2A25, OR2A5, OR2A1, OR2A42, OR2A7*	Olfactory receptor, family 2, subfamily A	...	RP11-703N5, RP11-466J6
8p23.3	5	*OR4F21, OR4F29*	Olfactory receptor, family 4, subfamily F	...	RP11-418D21
11q11	8	*OR4C6, OR4P4, OR4S2, OR5D13*	Olfactory receptor, family 4, subfamily C,P,S,D	...	RP11-626N6
11q12.3	3	*ROM1*	Retinal outer segment membrane protein 1	Retinitis pigmentosa, digenic	RP11-484M5
12p13.2	3	*TAS2R14, TAS2R44, TAS2R48, TAS2R49, TAS2R50*	Taste receptor, type 2, member 14,44,48,49,50	...	RP11-202N1
12q13.2	3	*OR6C2, OR6C4, OR6C68, OR6C70*	Olfactory receptor, family 6, subfamily C	...	RP11-222A15
14q11.2	61	*OR4M1, OR4Q3, OR4K1, OR4K2, OR4K5, OR4N2, OR4K13, OR4K14, OR4K15*	Olfactory receptor, family 4, subfamily M,Q,K,N	...	RP11-597A11, RP11-490A23, RP11-449I24, CTD-2024K23
15q11.2	26	*OR4M2, OR4N4*	Olfactory receptor, family 4, subfamily M,N	...	RP11-281J20
16p13.3	7	*OR1F1*	Olfactory receptor, family 1, subfamily F	...	RP11-680M24
17q25.3	18	*ACTG1, FSCN2*	Actin, gamma 1 propeptide; fascin 2	Deafness, autosomal dominant 20/26; retinitis pigmentosa-30	RP11-730A9, RP13-550B21
19p13.2	62	*OR2Z1*	Olfactory receptor, family 2, subfamily Z	...	RP11-282G19, RP11-367L15
22q11.1	15	*OR11H1*	Olfactory receptor, family 11, subfamily H	...	RP11-561P7
22q12.3	5	*MYH9*	Myosin, heavy polypeptide 9, nonmuscle	Deafness, autosomal dominant 17	RP11-108P21

[a] Total number of copy-number gains and losses observed for a CNV locus.

Now, consider what a pedigree looks like with significant CNV

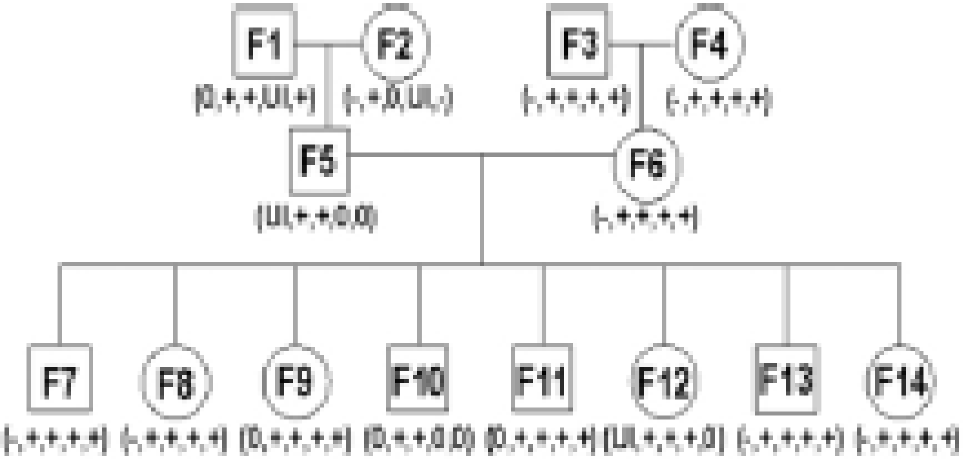
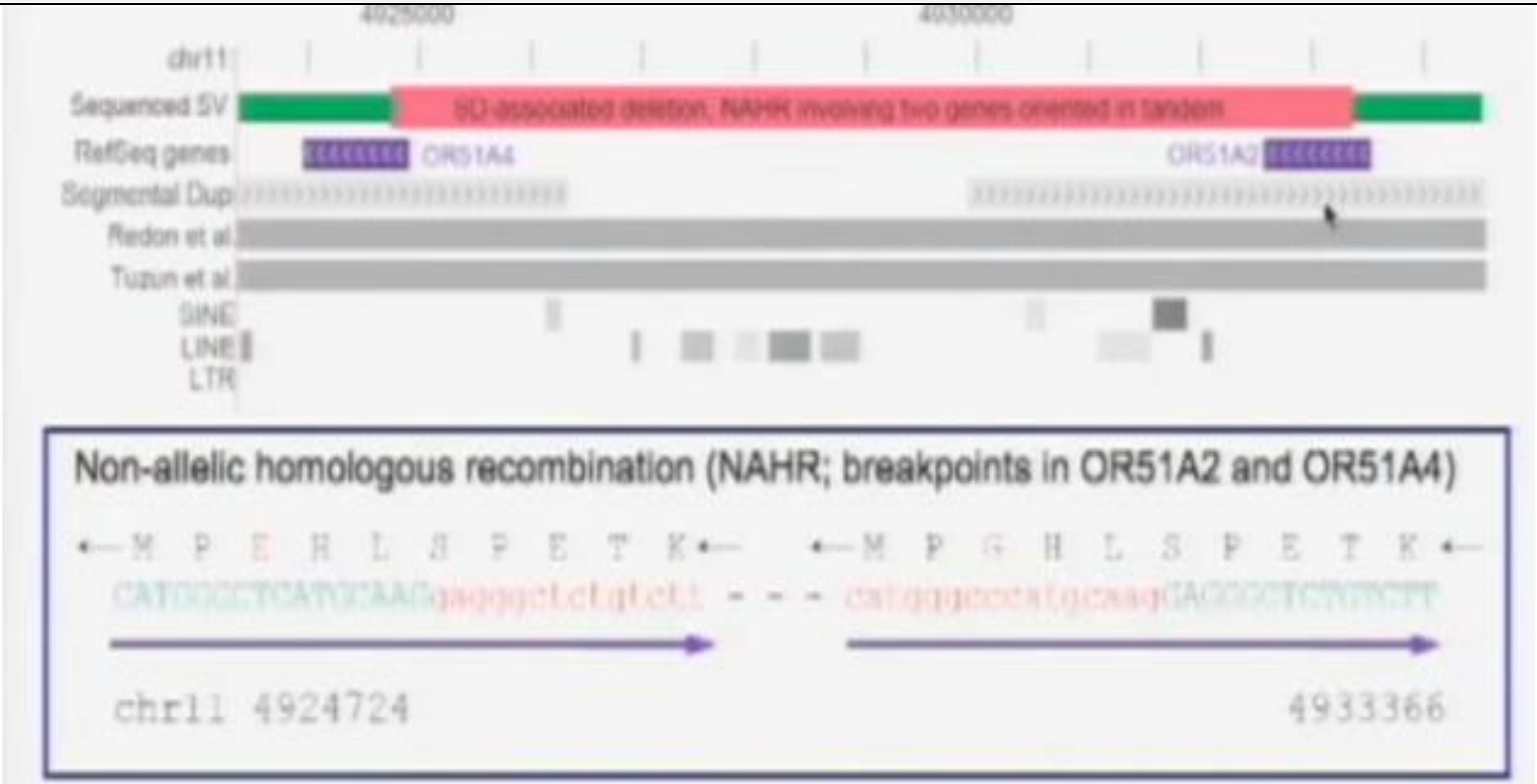


Figure 8. Inheritance of CNVs at five olfactory receptor loci in 14 members of a CEPH pedigree. The five loci (and clones), in the order shown, are *OR2A1* (RP11-466J6), *OR2Z1* (RP11-367L15 and RP11-282G19), *OR4K1* (RP11-449I24 and CTD-2024K23), *OR4M1* (RP11-597A11), and *OR4Q3* (RP11-490A23). – = Copy-number loss; + = copy-number gain; 0 = no copy-number change; UI = uninfomative. Male and female family members are shown as squares and circles, respectively.

Closer look at one of these Olfactory Receptor structural variations. Internal deletion of adjacent receptor genes creates fusion RNA/protein

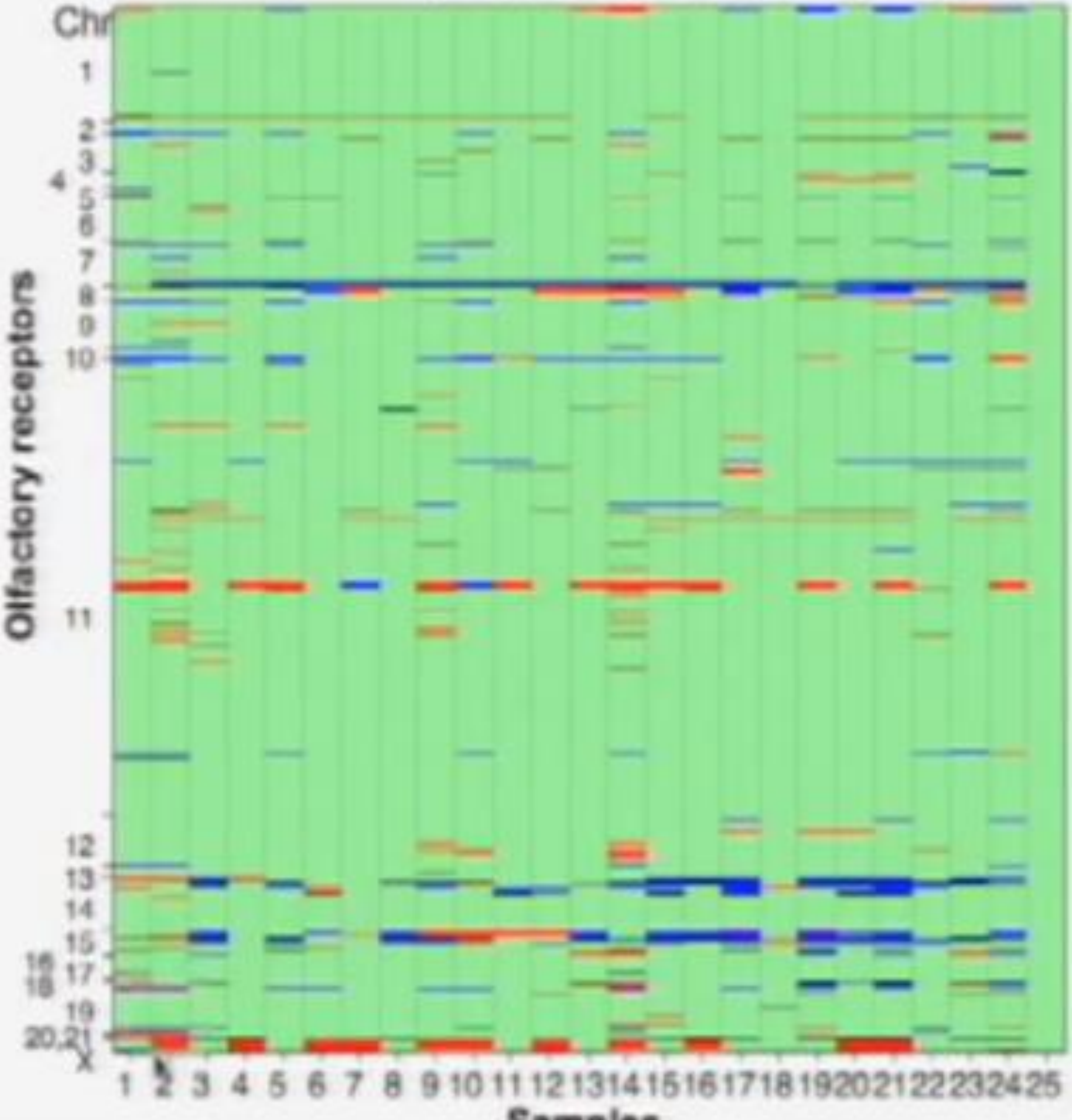


Olfactory Receptor Gene Fusion

graphic M. Snyder

Differences in Olfactory Receptor Genes

(Examined 851 OR Loci)

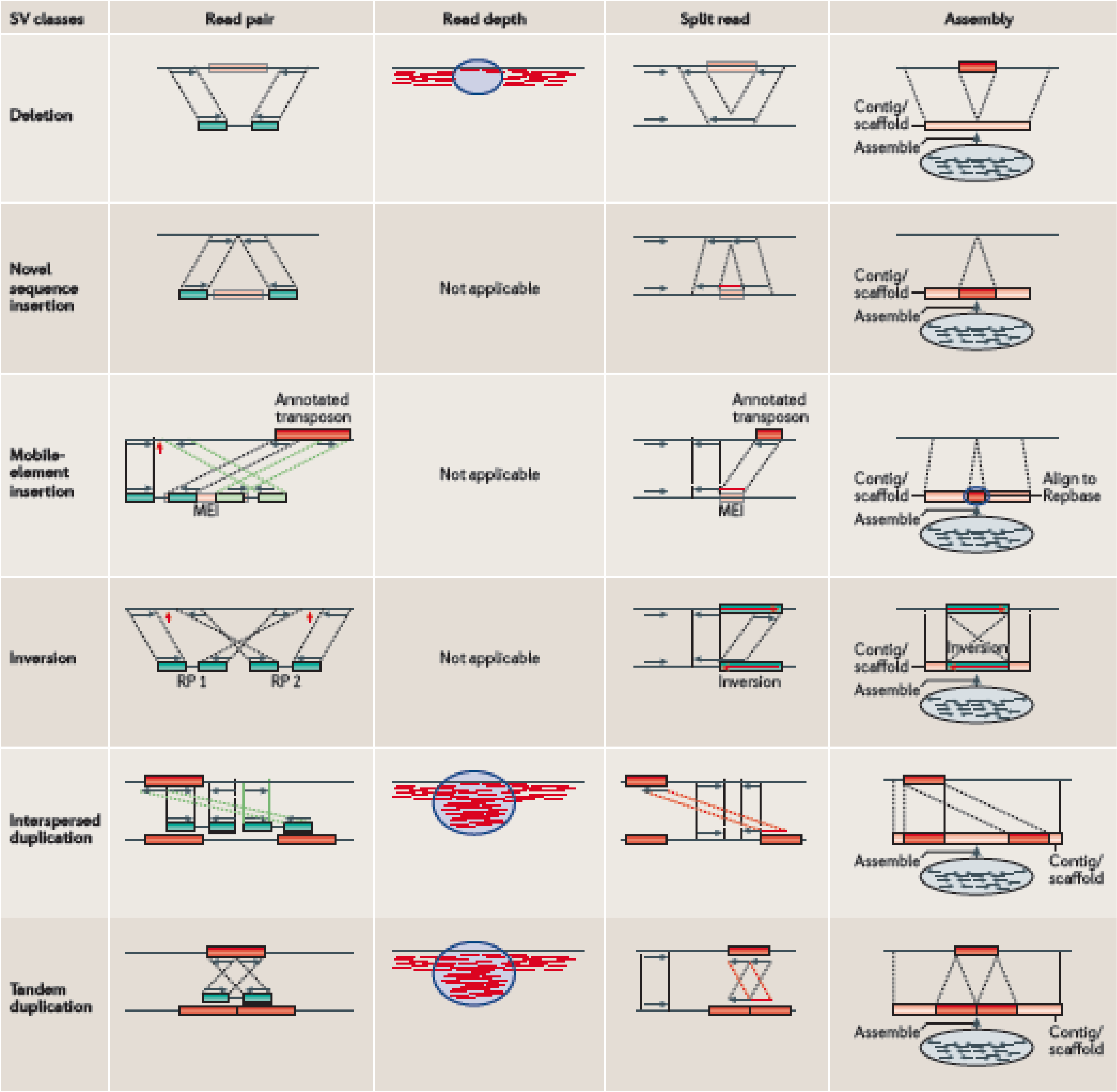


🟥 Gain

🟦 Loss

🟩 No change

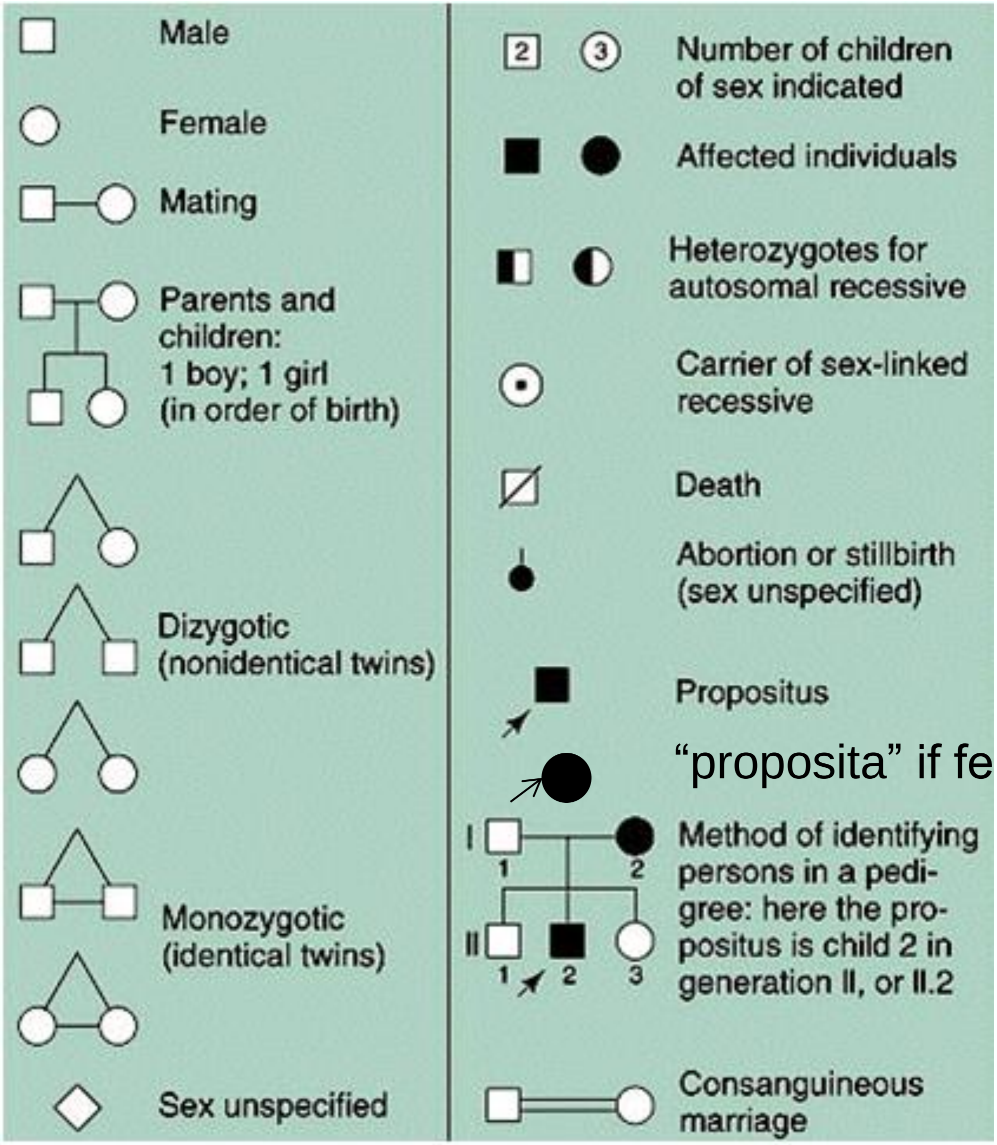
CNVs affect:
93 Genes
151 ψgenes



Consider how and what you can learn about each event class by direct modern sequencing

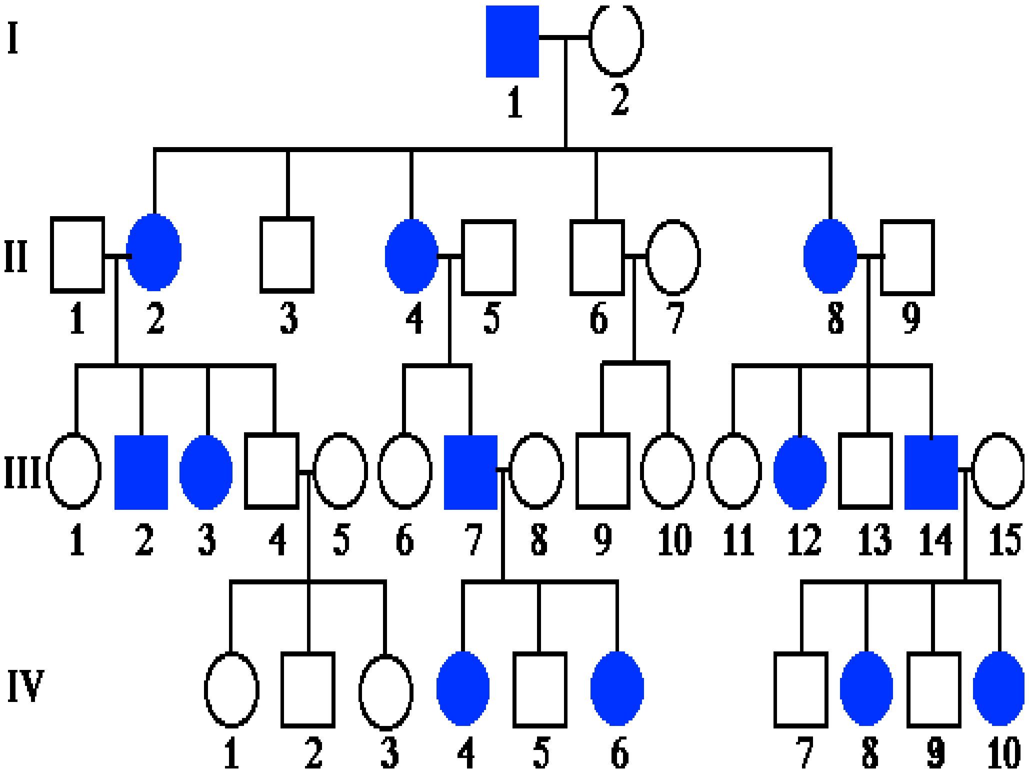
Broad intro to Short Read “Next Gen” DNA sequencing

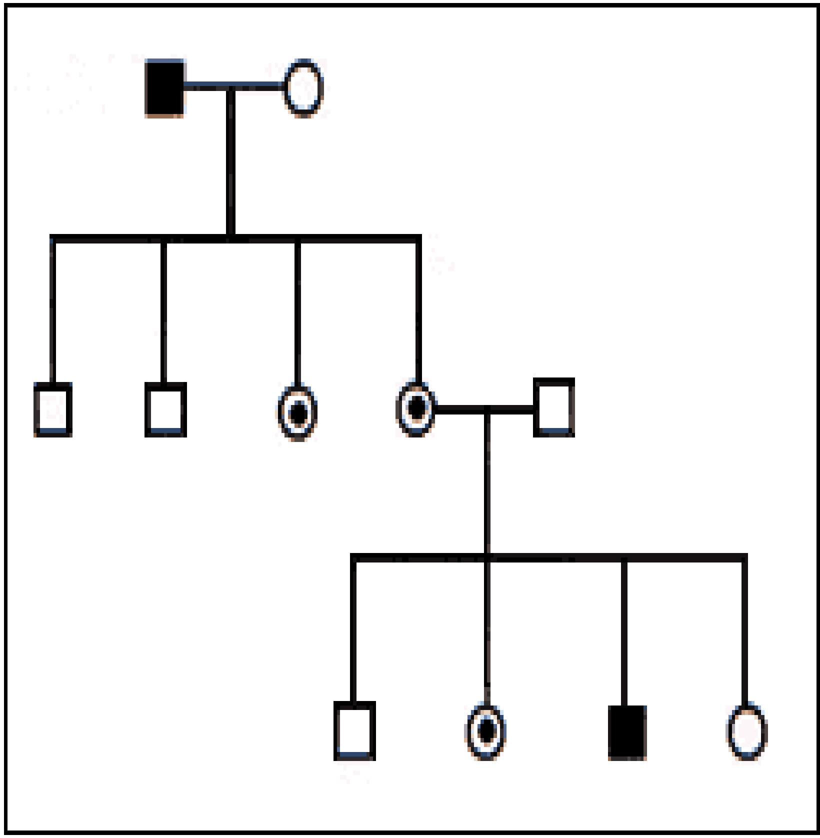
Human Pedigree graphic conventions


Male
Female
Mating
Parents and children: 1 boy; 1 girl (in order of birth)
Dizygotic (nonidentical twins)
Monozygotic (identical twins)
Sex unspecified
2
3
Number of children of sex indicated
Affected individuals
Heterozygotes for autosomal recessive
Carrier of sex-linked recessive
Death
Abortion or stillbirth (sex unspecified)
Propositus
I
II
1
2
1
2
3
Method of identifying persons in a pedigree: here the propositus is child 2 in generation II, or II.2
Consanguineous marriage

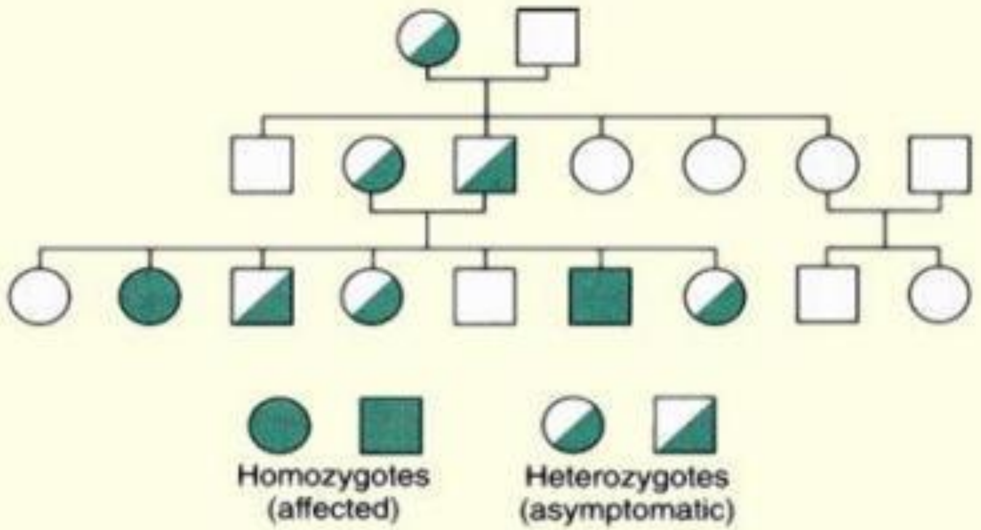
"proposita" if female

Idealized X-linked dominant and recessive pedigrees



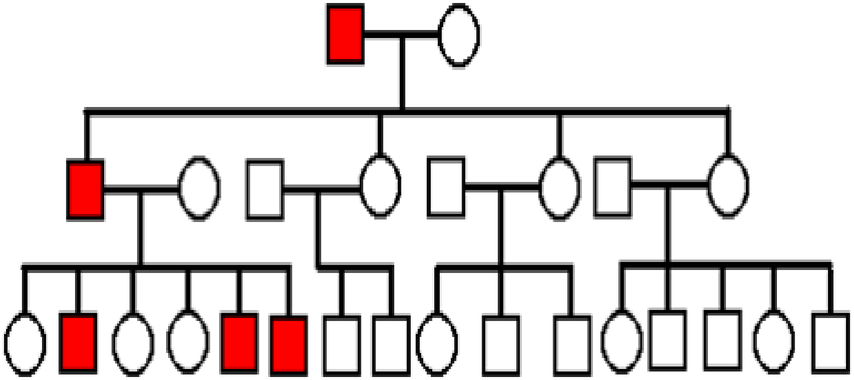


Autosomal Recessive Pedigree



Example you know: CF = cystic fibrosis CFTR gene

Y genes and Y-Linked inheritance



ASMTY (acetylserotonin methyltransferase),
TSPY (testis-specific protein),
IL3RAY (interleukin-3 receptor),
SRY (sex-determining region),
TDF (testis determining factor),
ZFY (zinc finger protein),
PRKY (protein kinase, Y-linked),
AMGL (amelogenin),
CSF2RY (granulocyte-macrophage, colony-stimulating factor receptor, alpha subunit on the Y chromosome),
ANT3Y (adenine nucleotide translocator-3 on the Y),
AZF2 (azoospermia factor 2),
BPY2 (basic protein on the Y chromosome),
AZF1 (azoospermia factor 1),
DAZ (Spermatogenes is deleted in azoospermia),
RBM1 (RNA binding motif protein, Y chromosome, family 1, member A1),
RBM2 (RNA binding motif protein 2), and
UTY (ubiquitously transcribed TPR gene on Y chromosome).
USP9Y
AMELY

Many occur in multiple copies with rich psuedogene representations. Prominent spermatogenesis functions, as expected

Intro for next time

- Exome – definition theoretical
- Operational definition
  - Concept of “expanded” Exome “conserveome”
- Relevance for finding rare mutations
  - Mendelian traits – especially monogenic
  - somatic mutations in coding sequence
    - importance of constraint from triplet code

  - Paper Ng et al. 2010 DHODH