Bi 188, 2011
meeting and lecture 1

Note that a slide has been added to answer questions

On our homework policy; course material distribution; expectations concerning literature citations in your paper. Any questions on any aspect of our policies will be answered at the beginning of meeting 2.

Reading in support of lecture 1 is given on slide 5.

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_
Georgi Marinov Georgi@caltech.edu_

Specific hours determined per doodle-poll [Monday and Tuesday are eligible

1. **The book and lectures are not highly redundant. The book is intended as background material for which you will be responsible. It is suggested that after the first week you complete the reading before coming to the lecture. Chapters 1-7 are for filling in and brushing up on relevant molecular biology.**

2. **Most lectures have additional reading from the literature. Generally this will include one or two review or summary pieces (which are best to read first) and one primary paper. With the exception of preprints, most papers will be from journals and you will download them using Web of Science etc.**

3. **There will be a midterm, a genome analysis project/paper, and a final exam.**

TAs Exams, optional home works, required paper

Katherine Fisher-Aylor x4923 kfisher@caltech.edu
Georgi Marinov x4923 georgi@caltech.edu

Office hours – 128 Kerckhoff finalized via doodle poll

Provide your email to the TAs if you are enrolled or auditing

30% midterm
30% final
40% paper/project

18% available as extra credit problems - 6 sets - 3% each
Generally due Friday, beginning of class (3:00pm) electronically.
First set due Fri 3:00 pm April 8.

Policy on notes, homeworks, use of class materials

It is a Caltech Honor Code violation for students to post BI 188 course materials online or to transmit them outside the Caltech community. Anyone at Caltech who receives BI188 course materials from you is similarly prohibited from posting them or delivering them to anyone outside the community of Caltech students. You are responsible for making this clear to anyone to whom you give materials.

The usual Caltech homework rules apply: Discussions among students are approved, as are discussions with Tas, but you must ultimately do the problem and generate the answers on your own. You must also write it up independently.

For your paper, *concepts* as well as facts and direct quotes must be referenced. I expect you to reference correctly and assiduously the primary published literature from academic journals, reviews, and books you have used. Website summaries like Wikipedia should be used sparingly or not at all (except to locate relevant primary references and the facts in them). Data portals such as the genome browser at UCSC <u>ARE</u> proper sources and they should be referenced as such. The OMIM site is a proper starting point and can be referenced, but you are expected to go beyond it into the published literature.

Reading to support lecture 1

http://www.nature.com/nrg/journal/vaop/ncurrent/full/nrg2958.html

A substantial review on structural variation in the human genome and how we learn about it.

2 part pdf – a light-reading summation of status of human genome by various authors

Sequencing big eukaryotic genomes: How it was first done, basics of structure learned

Human was project impetus - completed 2003 (draft 2001)

2 projects - One clone based hierarchical shotgun by public consortium\

- Multiple individuals contribute to aggregate assembly; one individual per BAC region

  Subsequent finishing to $<10^{-4}$ error rate multiple individuals

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

Second was the first mammalian whole-genome shotgun (WGS) done by Celera Inc. Largely historic interest

- no finishing
- 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; reduced for inbred models

Review gene types, functions and genome composition board plus stats below

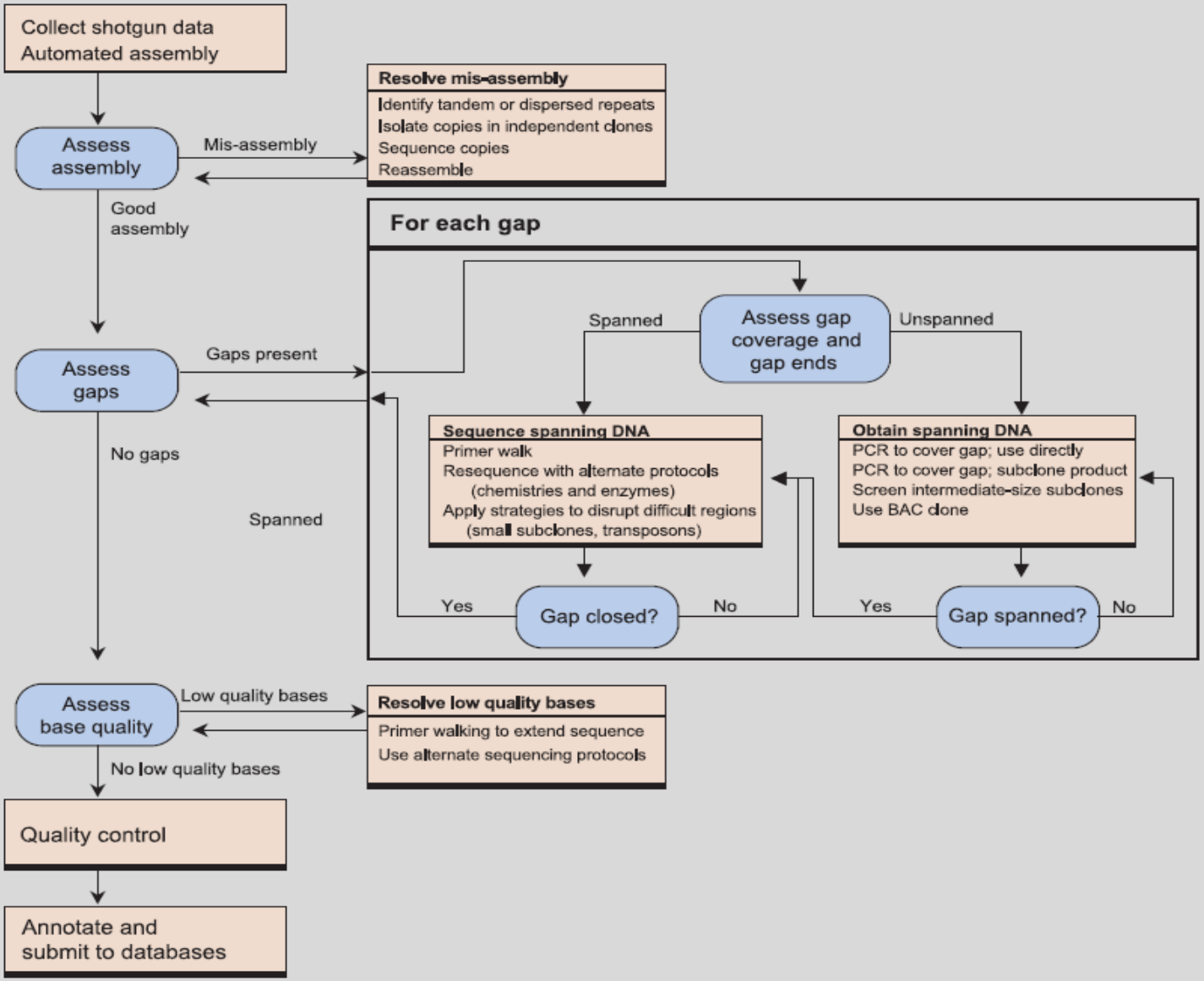
#BioType	Genes	Transcripts
IG_C_gene	16	18
IG_C_pseudogene	7	7
IG_D_gene	30	30
IG_J_gene	83	83
IG_J_pseudogene	3	3
IG_V_gene	180	181
IG_V_pseudogene	151	151
Mt_rRNA	2	2
Mt_tRNA	22	22
Mt_tRNA_pseudogene	580	580
TR_C_gene	3	3
TR_J_gene	13	13
TR_V_gene	48	48
TR_V_pseudogene	19	19
lincRNA	**1351**	**1592**
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

Note also pseudogenes;

Conceptual significance

Mechanisms of origin

Implications for various assays of gene expression



Box 2 Figure 1 Simplified flowchart for finishing of clones.

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

Chromosome	Euch. length* (bp)	N50† draft§ (bp)	Build 35 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.

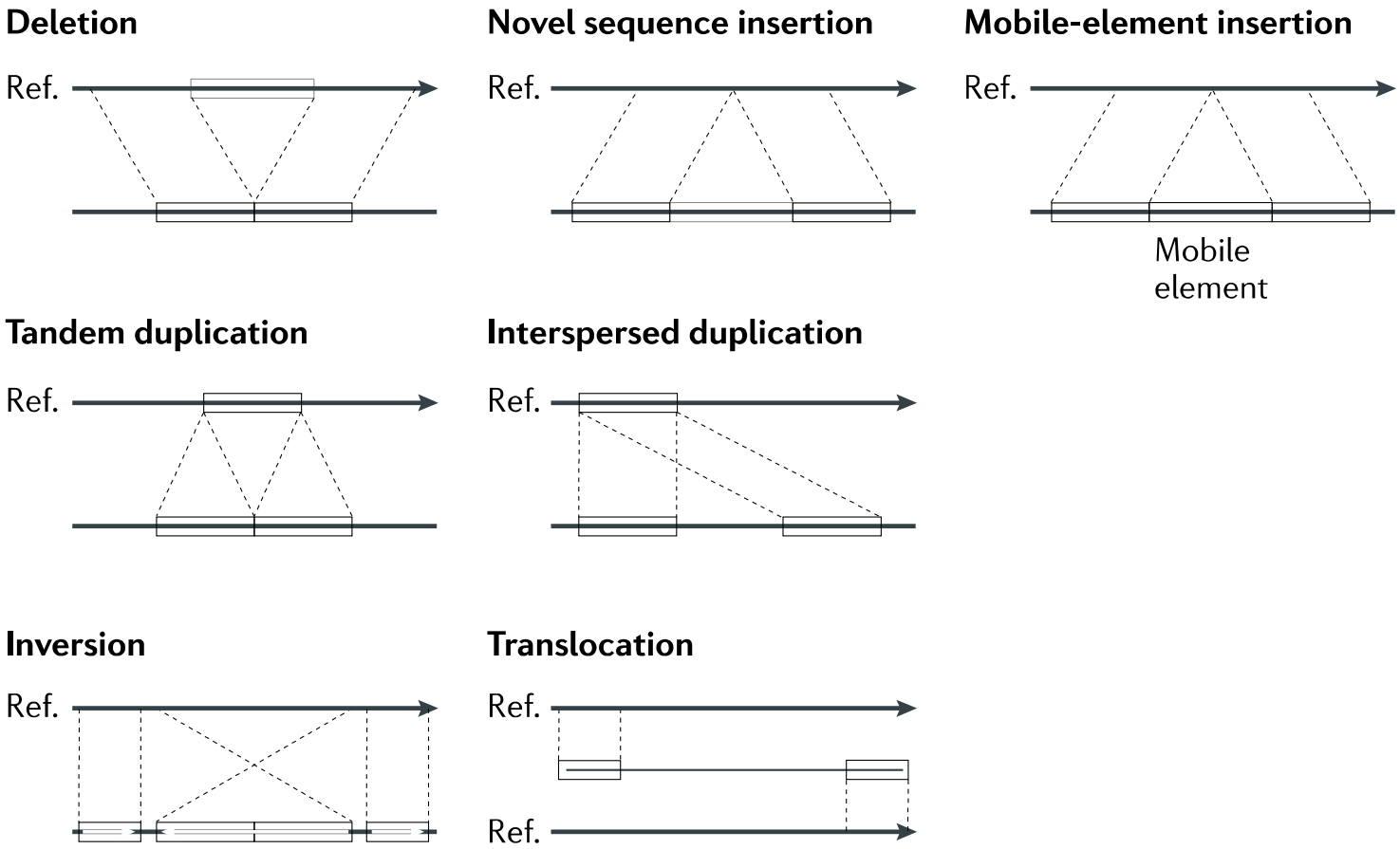
Useful metric: N50, which is the length in nucleotides at which 50% of the assembled genome is in blocks of the N50 size or longer

	panned Gaps		Unspanned Gaps			
chr	**All Scaffolds**	**Placed Scaffolds**	**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:

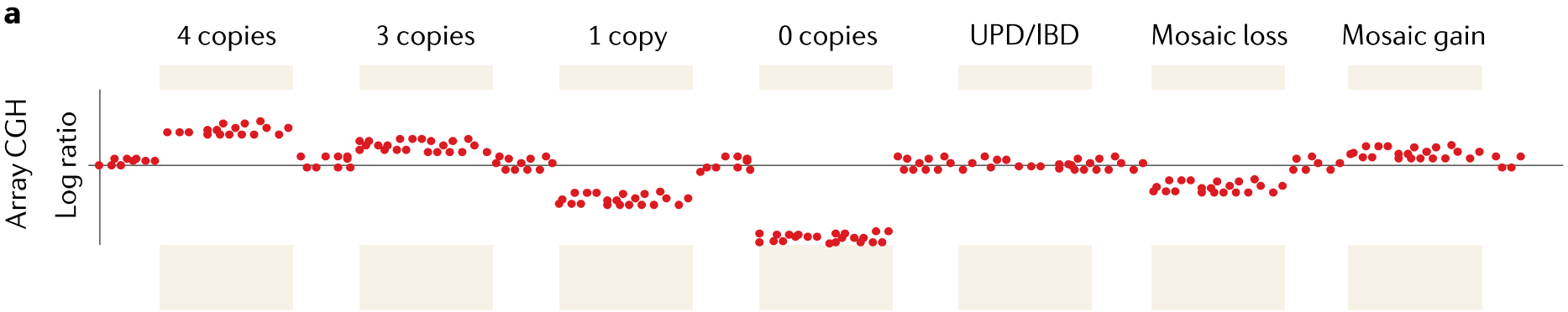
Distribution of GAPs in Current build of the human Genome

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



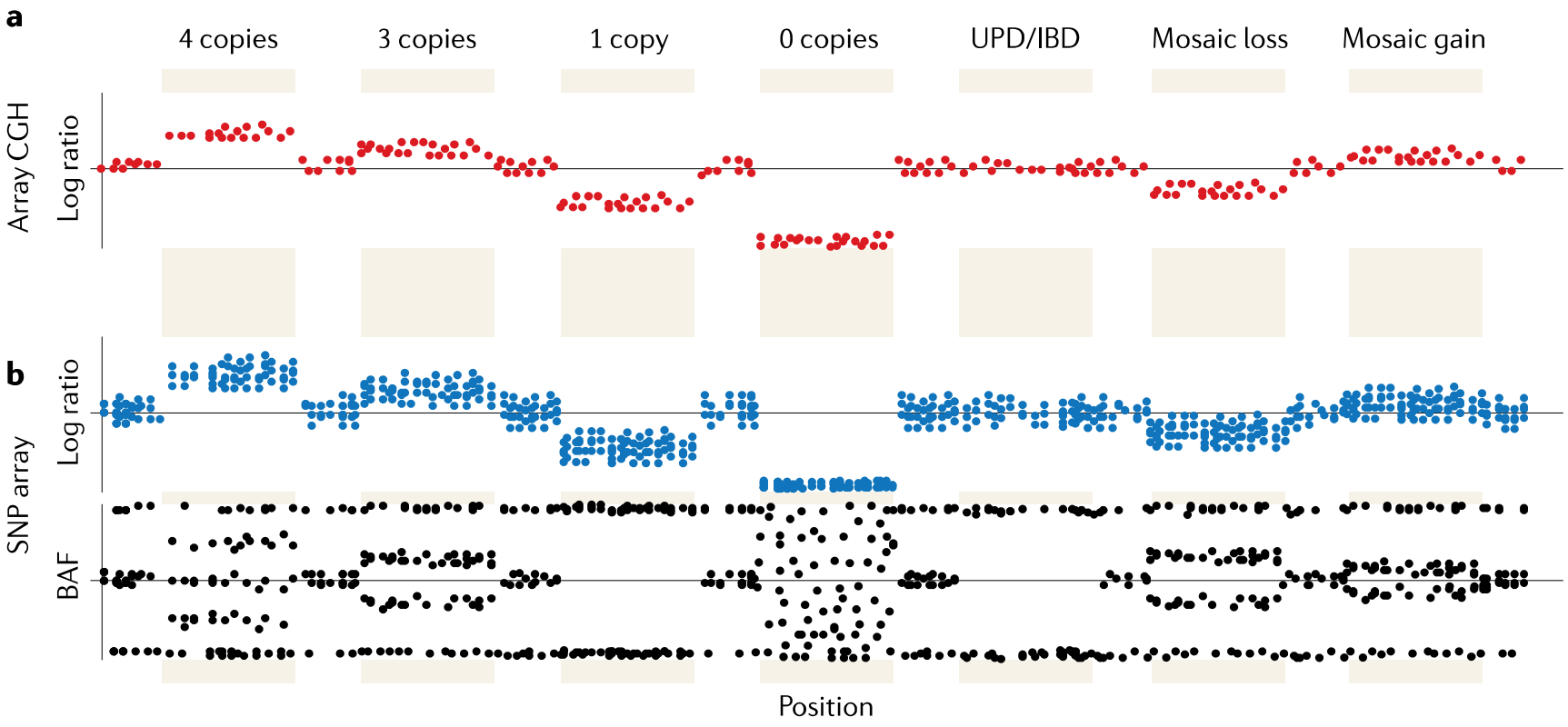
Copy number variation is a common and important consequence = CNV = differences in number for a gene or other sequence

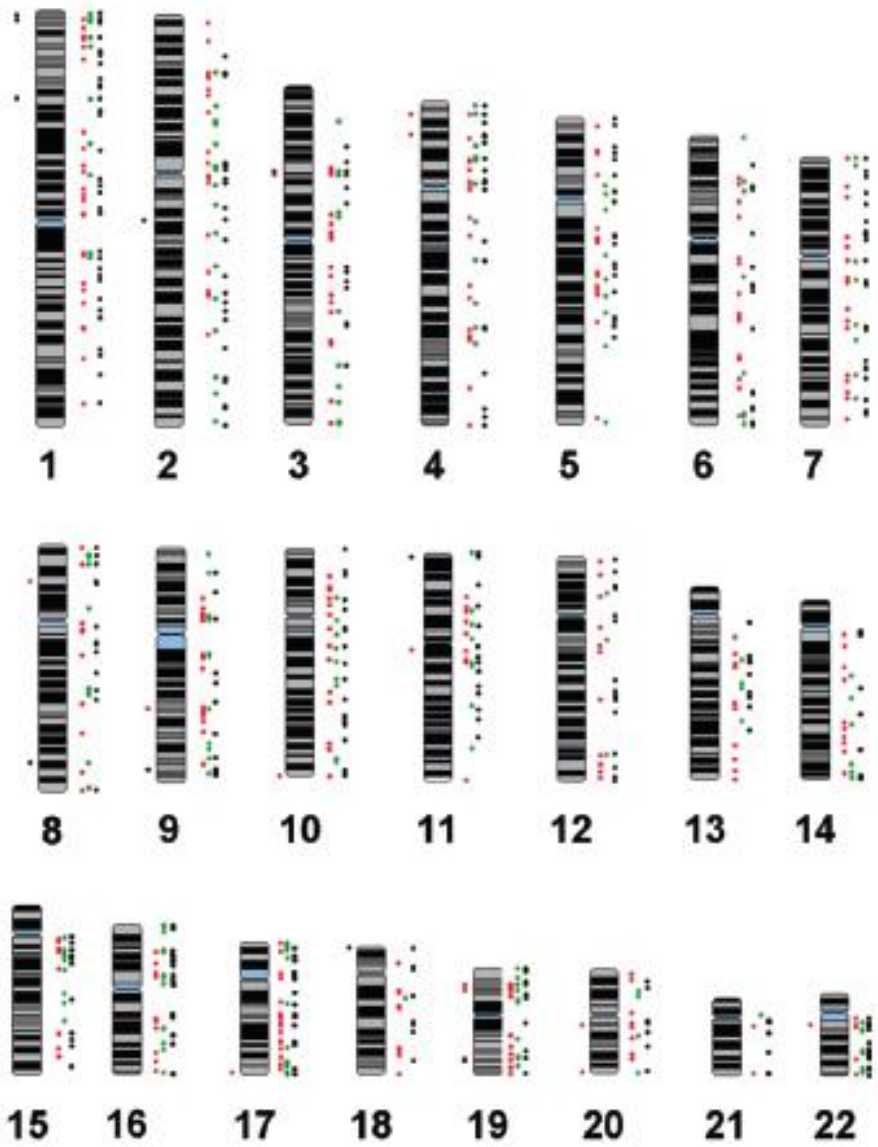
How is CNV detected experimentally? Multiple ways by now – differing issues of sensitivity, noise, resolution



Evan Eichler and colleagues; data via microarray CGH
[Review array hybridization principles – log 2 ratio probe a/b]

Compare CGH with SNP (single nucleotide polymorphism) array data





igure 6. Distribution of CNV clones. High-frequency CNV clones are shown as dots to the right of each chromosome; red, green, and ack 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*).

Human Segmental Duplication Map

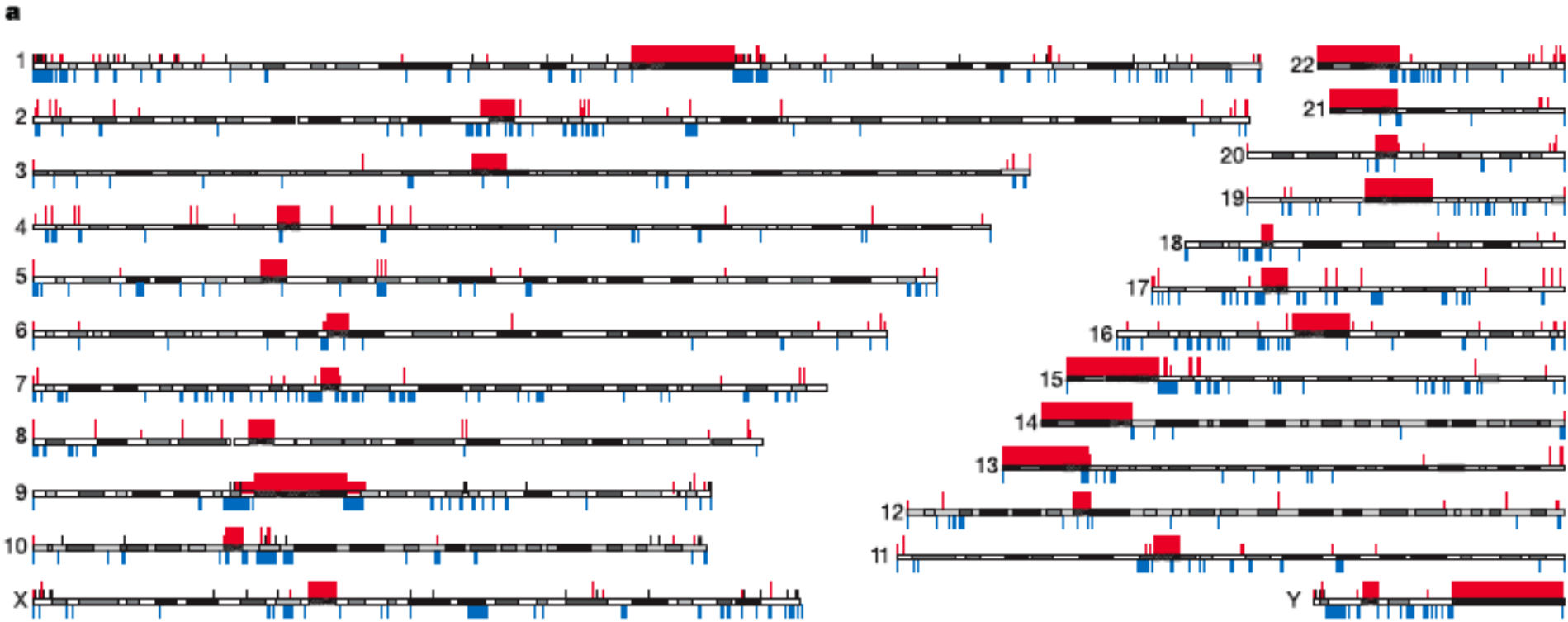
implications – functional and technical - for individual genomics

- 1kb to 500kb size
>90% similar

- 2 - 6 copies (up to 20)

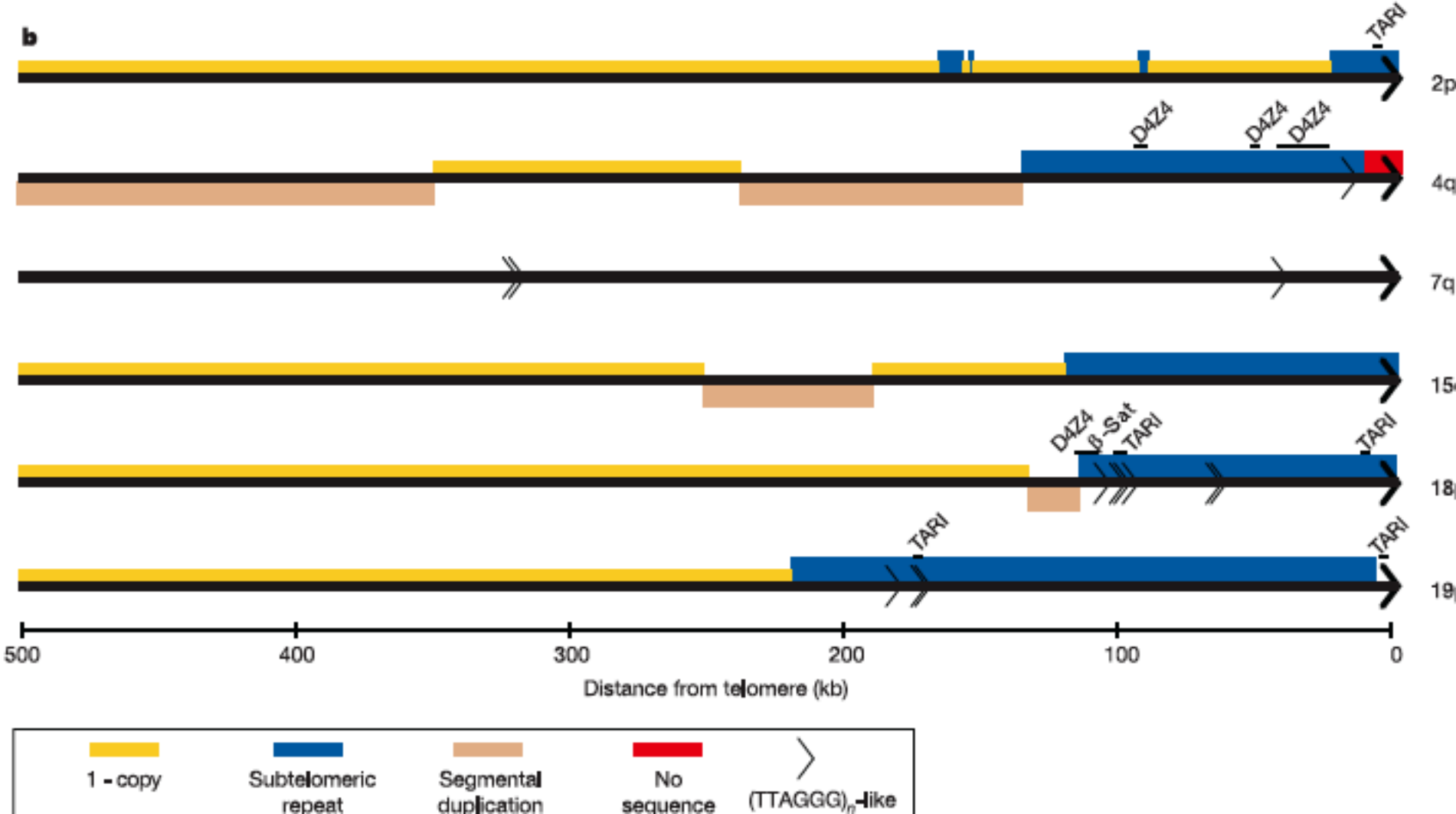
>5% of genome

Overall map shows range of sizes; telomeric representataion



Near telomeres > specialized repeats (blue) then Segmental dup events; then single copy

Sequence instability together with position effects (damping expression near sub-telomeric repeats).



Candidate biological significance groups
consider 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	40	*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	RP11-460M21, RP11-01A18
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	RP11-350D11
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	RP11-86D5
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, RP11-80B34
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.

Consider what a pedigree looks like with significant CNV

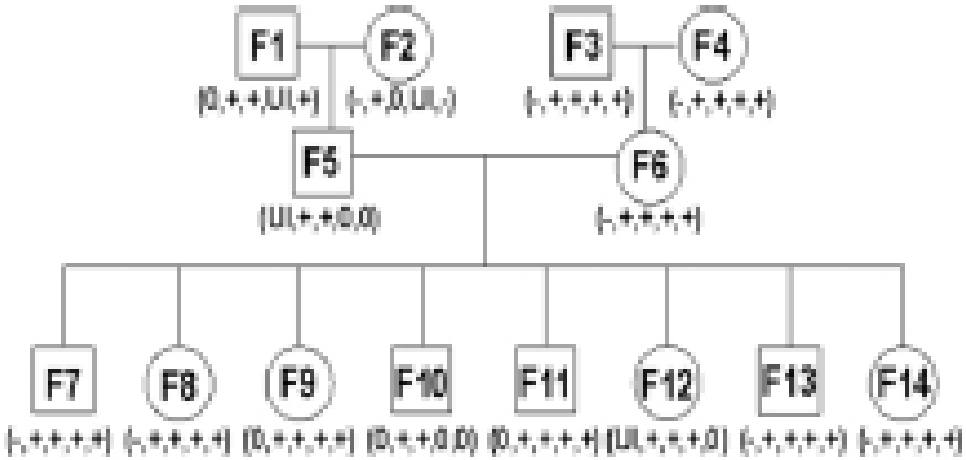
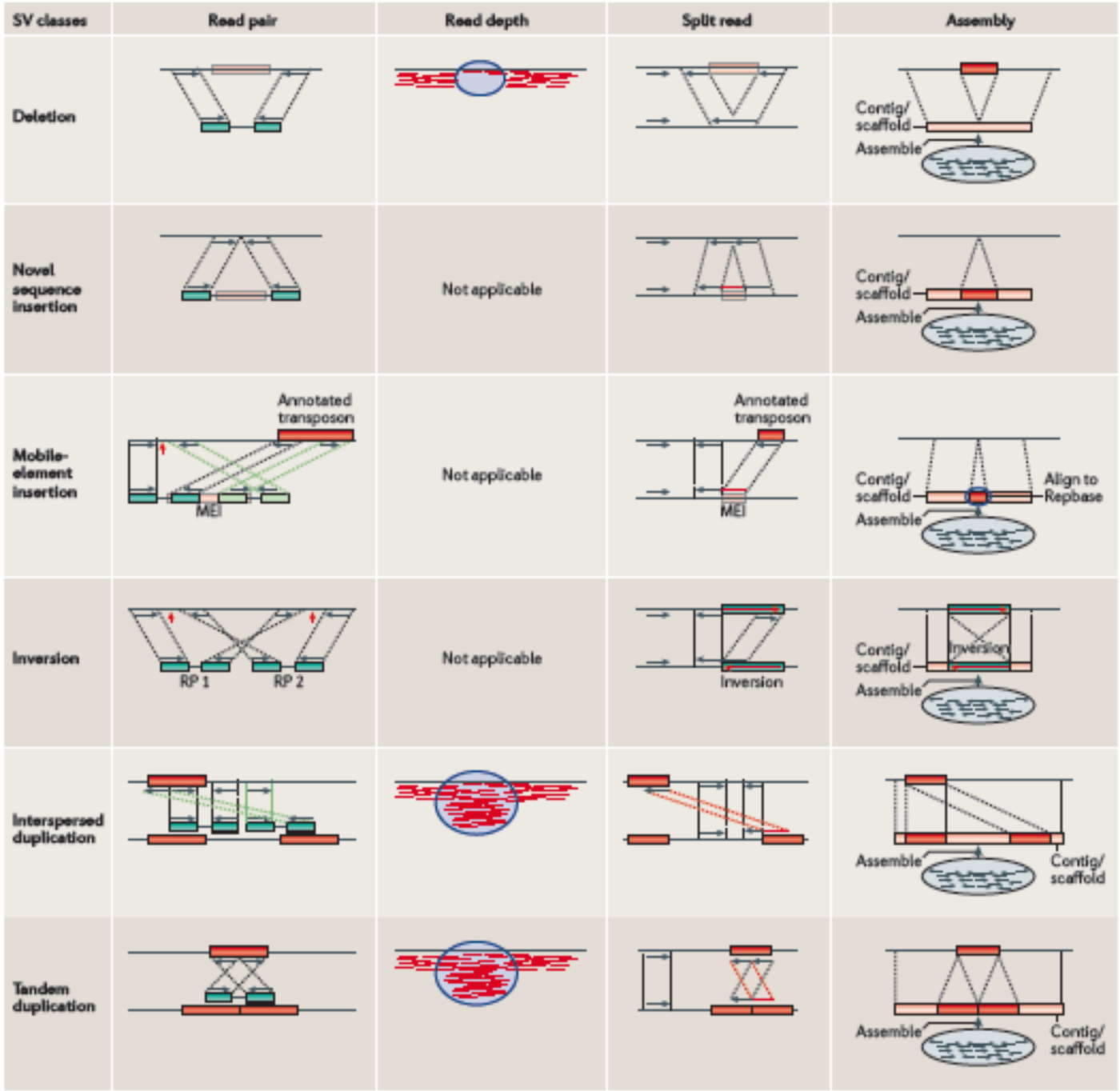


Figure 8. Inheritance of CNVs at five olfactory receptor loci in 14 members of a CEPH pedigree. The five loci (and clones), in the order shown, are *OR2A1* (RP11-466J6), *OR2Z1* (RP11-367L15 and RP11-282G19), *OR4K1* (RP11-449I24 and CTD-2024K23), *OR4M1* (RP11-597A11), and *OR4Q3* (RP11-490A23). − =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



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

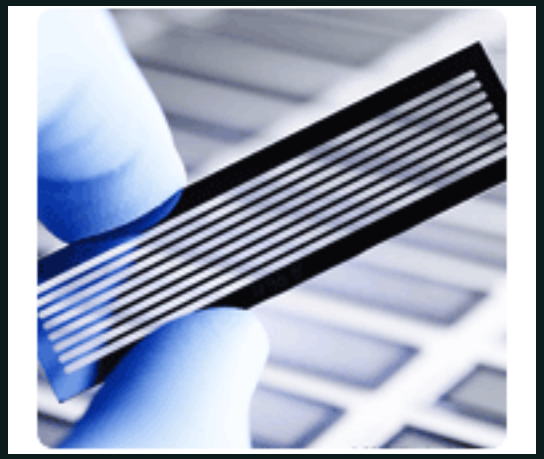
Board intro to Short Read “Next Gen” DNA sequencing

Technology often rate-limiting

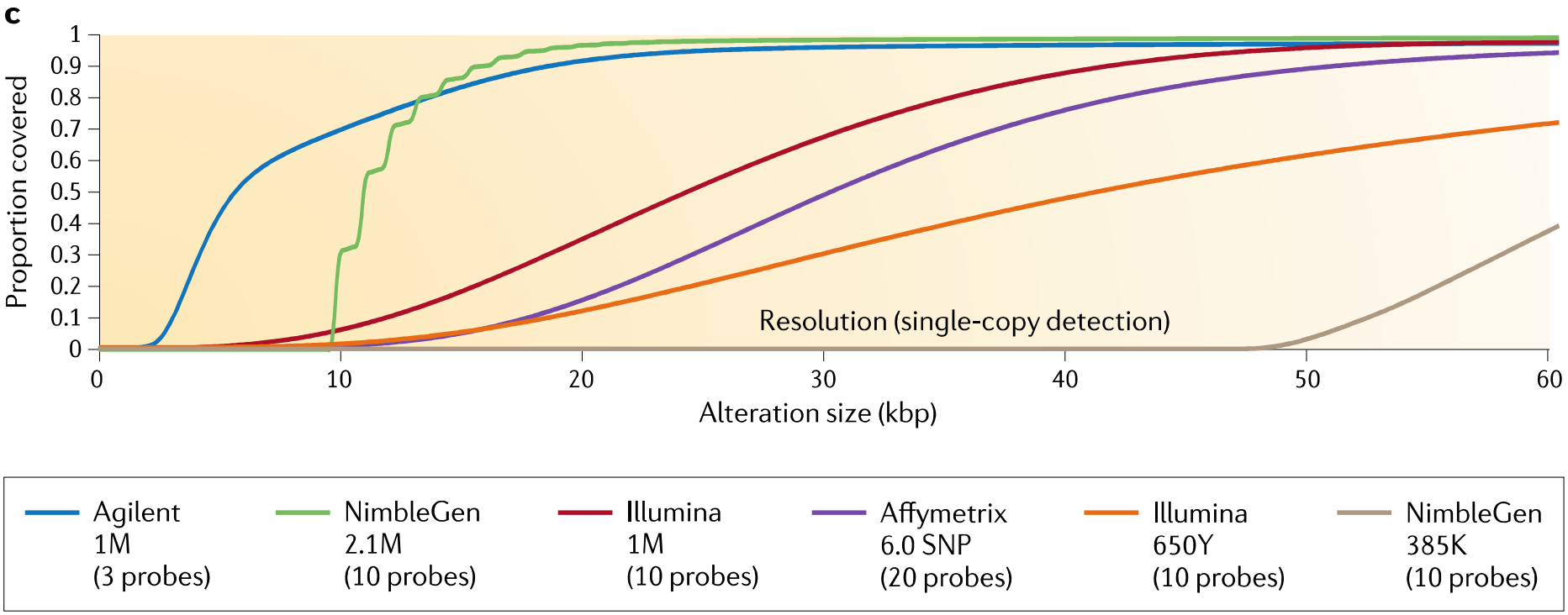
1998 - Audacious goal for DNA sequencing
2 million bases/ year/ entire U.S. Project

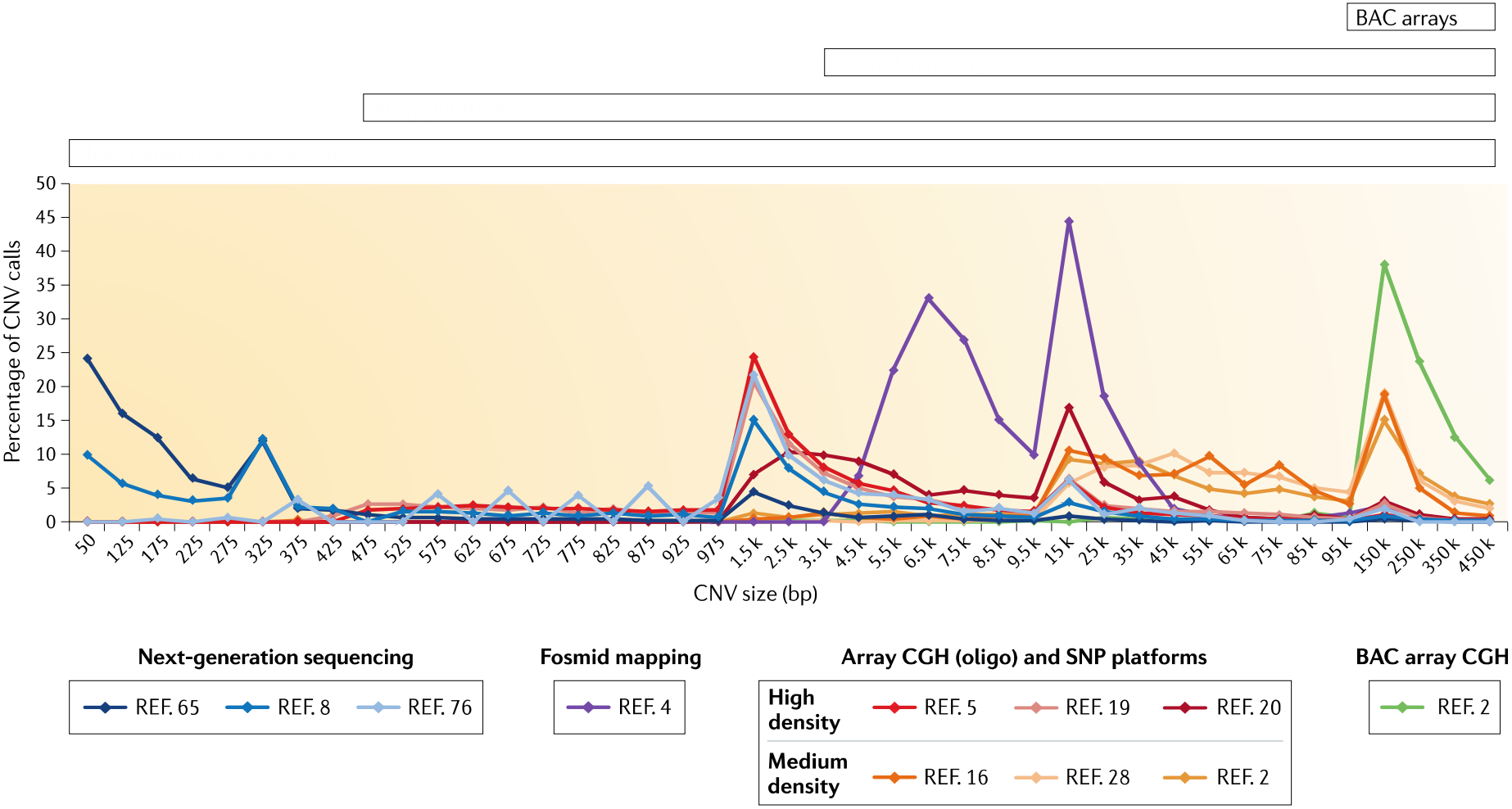
2009 - 2-4 billion bases/ 3 days/ machine

2011 - 200 billion - 1 terabases / 6 days / machine



As expected, the specifics of different array types
Affect resolution and sensitivity of an array CGH expt.




BAC arrays
Percentage of CNV calls
50
45
40
35
30
25
20
15
10
5
0
50
125
175
225
275
325
375
425
475
525
575
625
675
725
775
825
875
925
975
1.5 k
2.5 k
3.5 k
4.5 k
5.5 k
6.5 k
7.5 k
8.5 k
9.5 k
15 k
25 k
35 k
45 k
55 k
65 k
75 k
85 k
95 k
150 k
250 k
350 k
450 k
CNV size (bp)
Next-generation sequencing
REF. 65
REF. 8
REF. 76
Fosmid mapping
REF. 4
Array CGH (oligo) and SNP platforms
High density
REF. 5
REF. 19
REF. 20
Medium density
REF. 16
REF. 28
REF. 2
BAC array CGH
REF. 2