COMP90016 Computational Genomics: Structural Variations in DNA and Bioinformatics Detection Methods

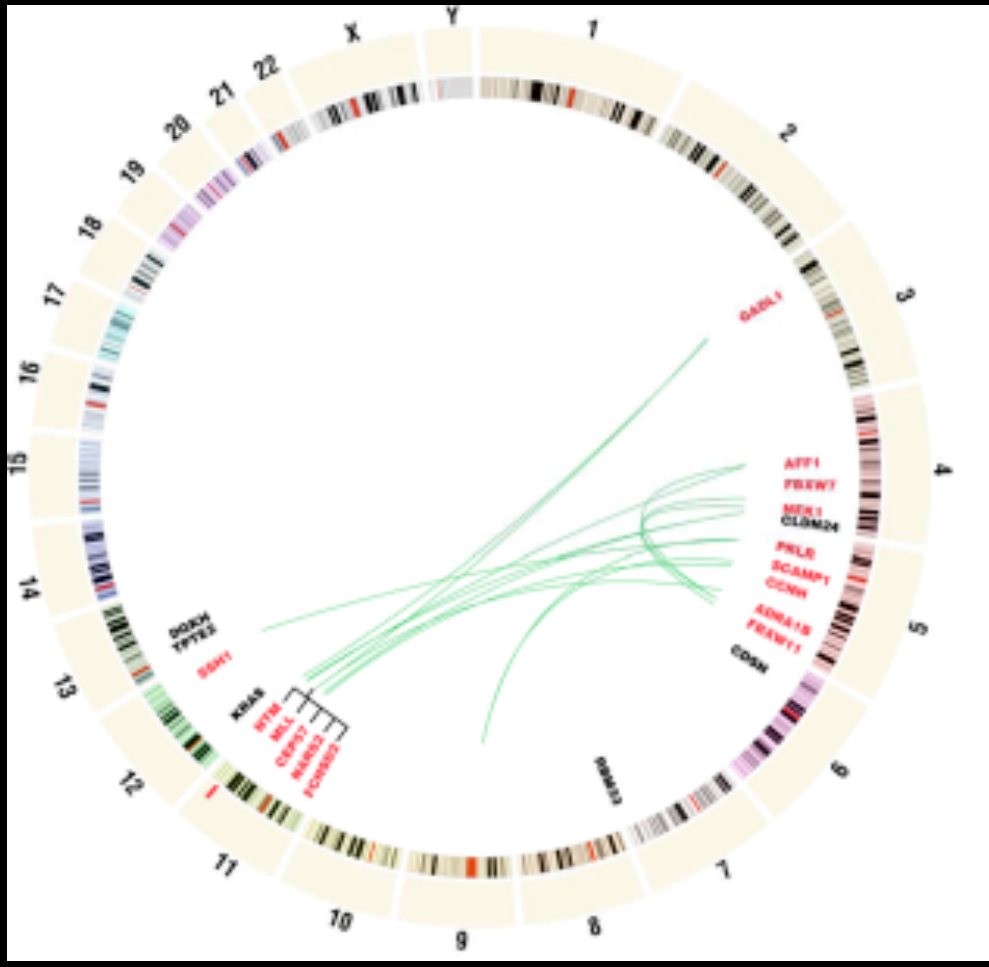


Image from: Dobbins et al 2013: **silent mutational landscape of infant *MLL-AF4* pro-B acute lymphoblastic leukemia**

Aim of this lecture

- Introduce structural variations/variants (SVs)
- Look at the signal created by different SVs in various types of data.
- Look at methods to detect some of the SV sub-types.

What are Structural Variations?

- Blurry lines of definition.
  - Genomic rearrangements between two sets of DNA.
  - Anything that is not a SNP?
  - Almost… usually anything that is not a SNP or small indel.
- Some of the sub-categories include:
  - CNVs
  - Fusion events (translocations, inversions, …)
  - Transposable elements
- Each of these events can be 10s to millions of nucleotides in size!

Why do we care about SVs?

- They are harder to find, but just as important as SNPs.
  - They can change the ability of cells to access or transcribe genes.
- They occur frequently in our genomes – just as SNPs.
- Even more important in genetic diseases or conditions:
  - Cancer
  - Down Syndrome
- Being able to detect SVs can help to understand differences in phenotypes and what drives certain diseases.

Terms and Clarifications

- Donor Genome:
  - The DNA underlying an analysed sample.
  - This is the unknown of the experiment.
- Reference Genome:
  - A representative sequence of the population.
  - For example “the human genome”.
- Somatic events:
  - Variation that is present in diseased samples or tissues only.
  - For example variation that are present in tumour DNA, but not the normal tissue of a donor genome.

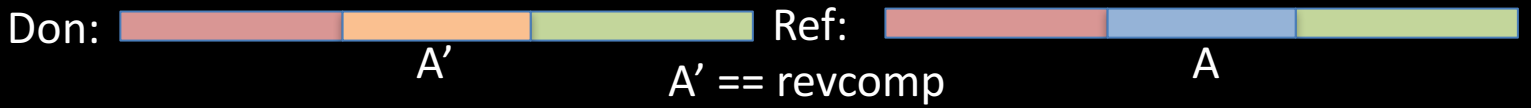
SV Examples

- Deletions:
  - Chunks of DNA that are present in the reference, but not in the donor genome.

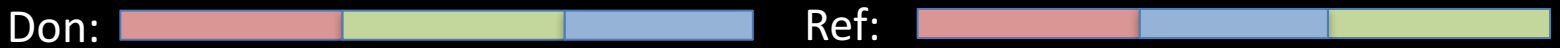
- Duplications:
  - Chunks of DNA that are present at more places in the donor than in the reference

SV Examples (2)

- Inversions:
  - Chunks of DNA which strandedness has been inverted in the donor with respect to the reference.



- Translocations:
  - Chunks of DNA that have moved to a different location in the donor genome.



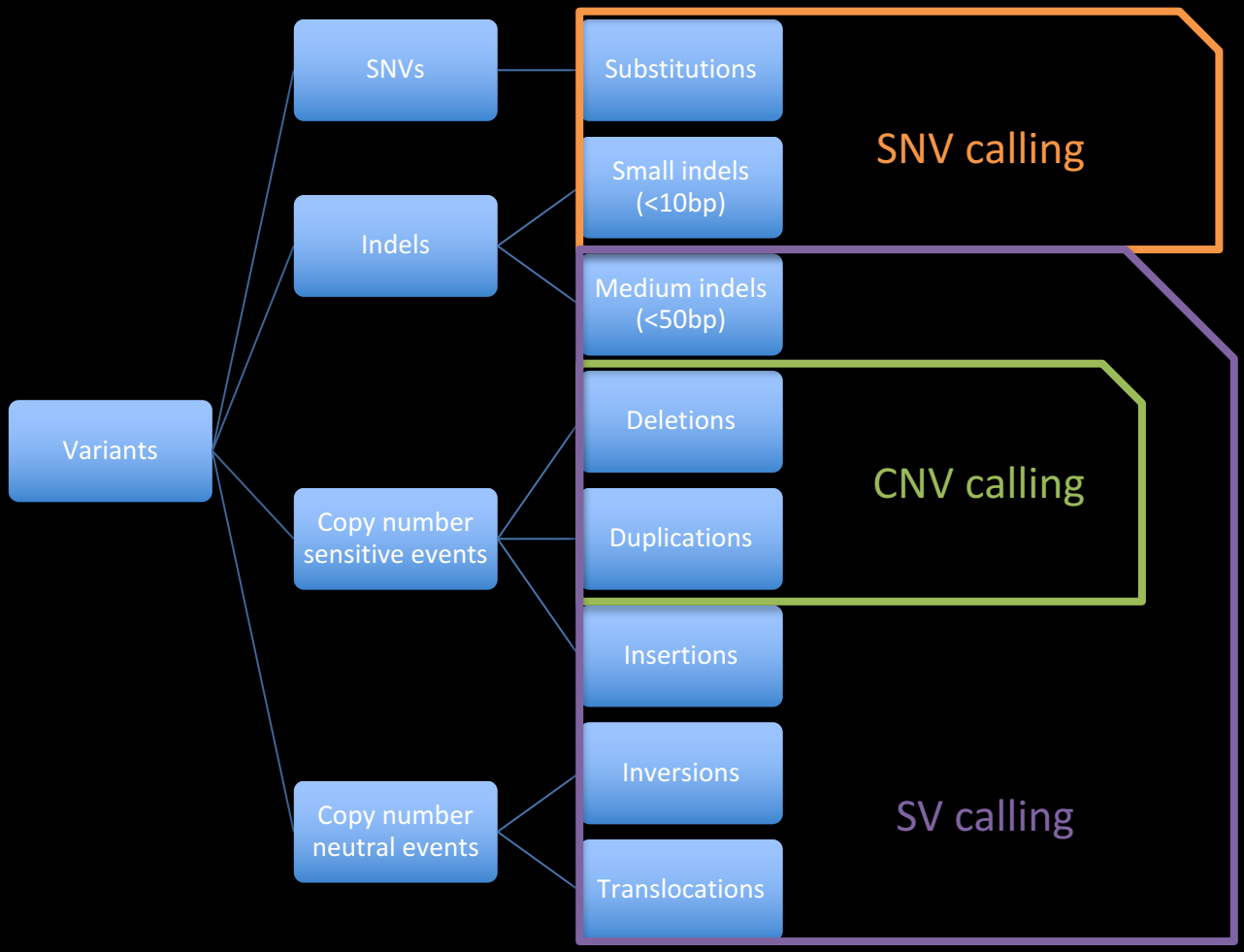
SV Examples (3)

- CNV == Copy Number Variation
  - Genomic rearrangements that reduce or increase the abundance of DNA at a location.
  - Sounds familiar? Same as Deletions and Duplications. Usually large in size (kilo- to megabases)
  - Different Techniques than for the other rearrangements are utilised to detect this subset.

Where Do SVs occur?

- Just as SNPs everywhere in the genome
  - "In total, we identified 277,243 SVs ranging in length from 1–23 kb." (Analysis of an Asian genome and an African genome by Li et al. in Nature: doi:10.1038/nbt.1904)
- A driving force in many cancer types:
  - Amplifying/altering/removing genes through structural variations enables a tumour to circumvent the cell's usual protection mechanisms.

Variations Overview



Data for SV analysis

- SNP arrays (as seen before)
- Fish
- Sequencing data
  - DNA seq (the focus of this lecture)
  - RNA seq
  - Exome seq

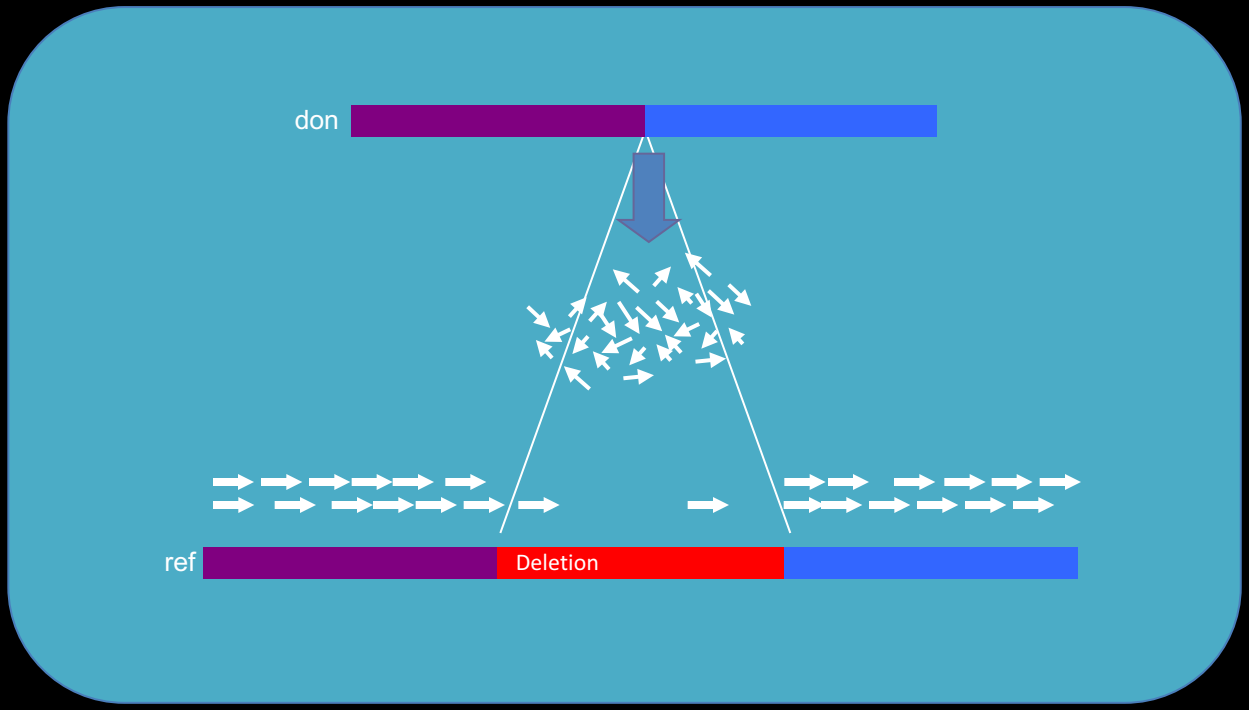
Sequencing Data

- Much more Versatile than traditional technologies – can detect all the types mentioned above
  - With higher precision
  - With more context (which bits are fused to each other etc)
- In sequencing data we can utilise four basic strategies:
  - Anomalously paired reads
  - Split-reads (Soft-clips)
  - Read Depth
  - De novo assembly

Read Depth Detection of CNVs

- Read depth == coverage == how many reads overlap a given location in the (reference) genome?
- What kind of SVs leave signal in read depth?
  - Deletions
  - Duplications (CNVs)

Depth-of-coverage



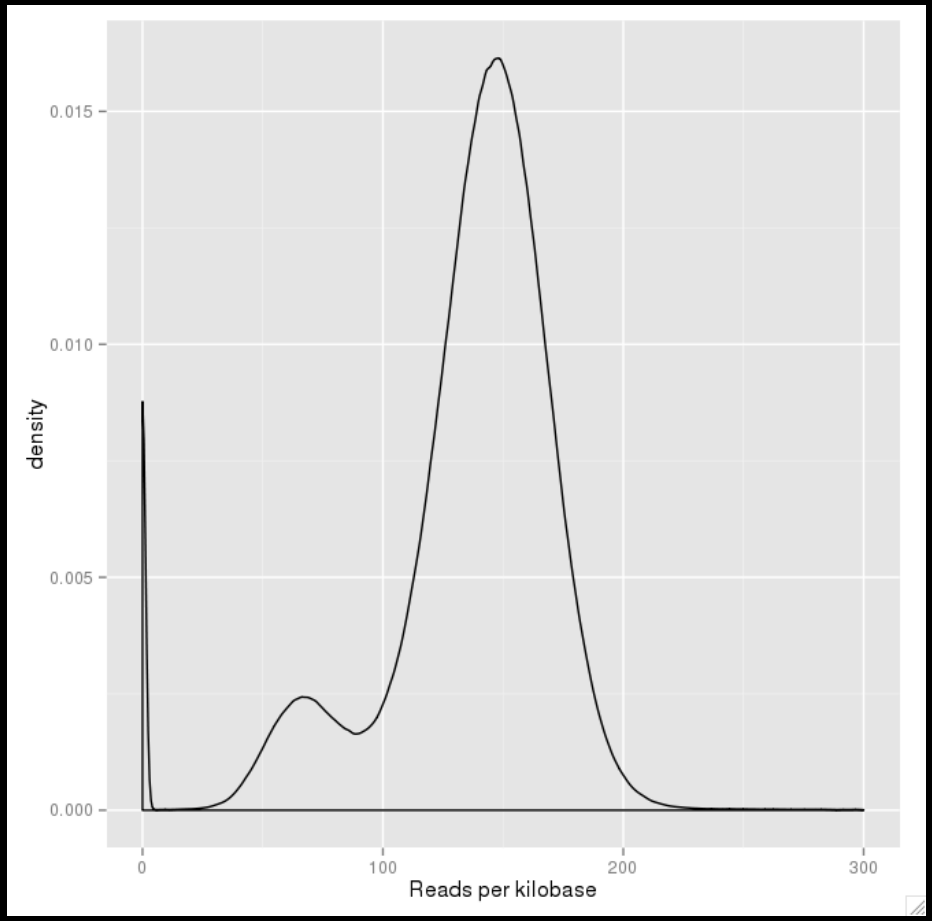
Depth-of-coverage can help detect SVs

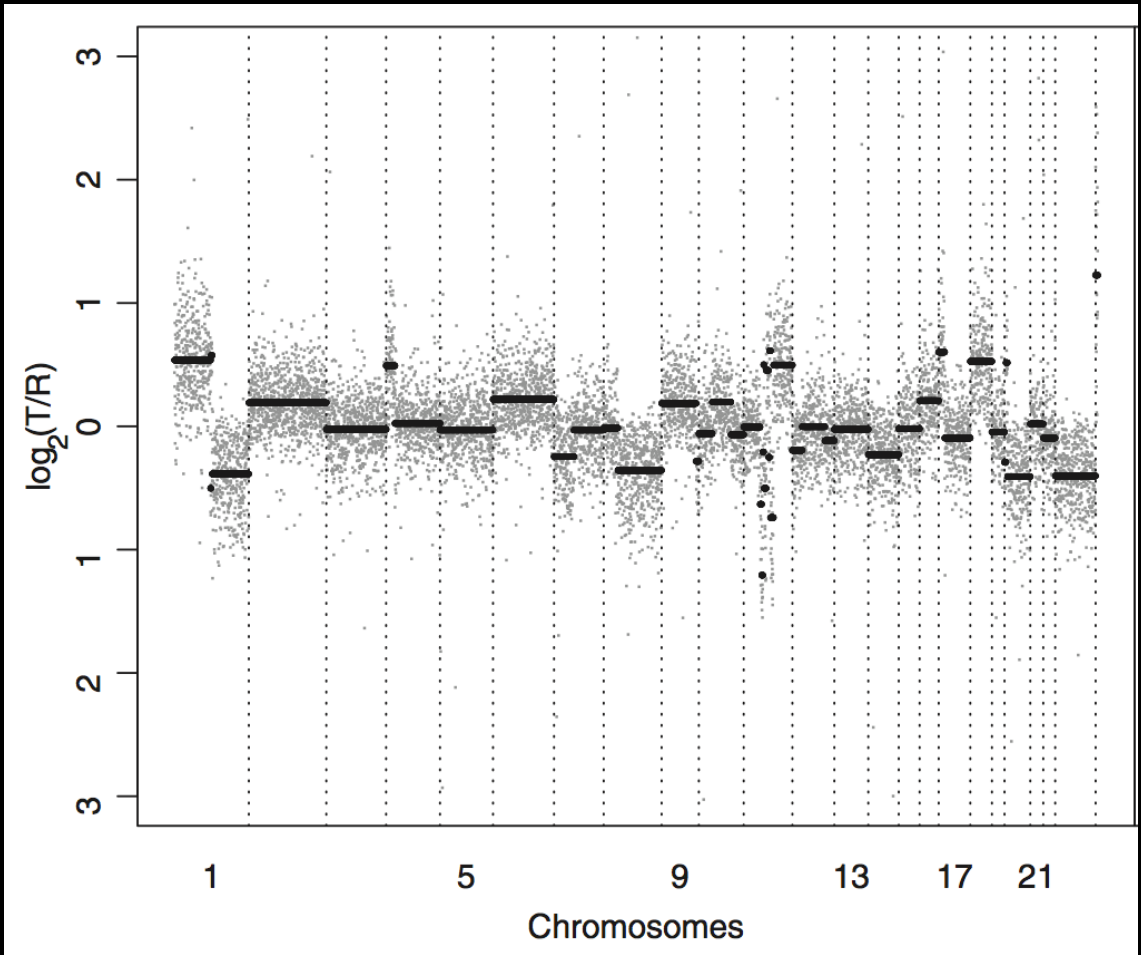
How Do Read Depth Methods Work in Practice?

- Binning strategy to control computational burden and variance of signal:
  - Chop the genome into non-overlapping segments (bins) of equal length (100bp, 1kbp, 5kbp, …).
  - Count reads (or bases) in each bin.
  - Analyse the distribution of bin sizes (what is the average, what is the variance?).
  - Identify bins that are extraordinary in size.
  - Merge neighbouring bins of similar size into copy number intervals.

Example for Distribution of Bin Sizes in Real Data

- A fairly well defined signal.
- Additional peaks for alternative copy number states.
- Rough copy number calling strategy:
  - Bins with 100-200 reads are normal (2 copies).
  - A lower count means loss of copy, higher count gain.





Circular binary segmentation for the analysis of array-based DNA copy number data (Olshen et al, 2004)

Circular Binary Segmentation (CBS)

- The idea of this algorithm is based on change points:
  - A change point is a point in the data that tested positively for significant change.
  - In this case the change is going to be different copy number states.
- Bin the data and let the log ratios of each bin be represented by (X1, X2, …, Xn).
- Let Si denote the cumulative sum of bins: Si=X1+X2+…+Xi
- To test for a change at i we analyse the following statistic:
  - $Z_i = \left(\frac{1}{i} + \frac{1}{n-i}\right)^{-1/2} \times \left(\frac{S_i}{i} - \frac{S_n - S_i}{n-i}\right)$

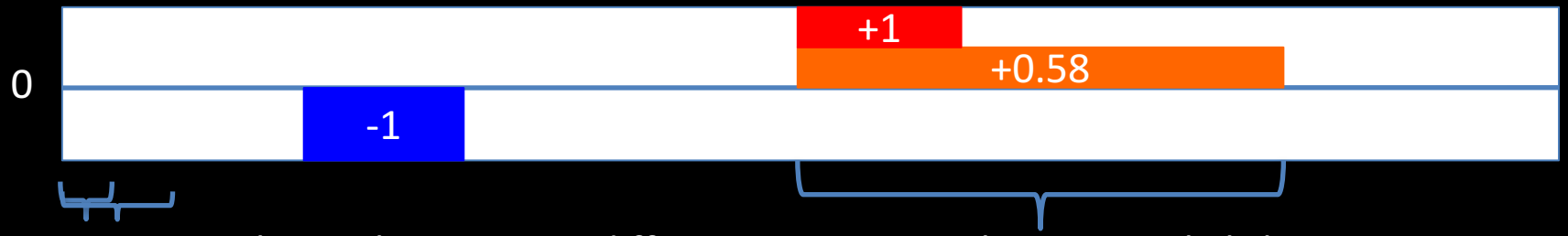
CBS continued

- The procedure above allows us to test for single change in copy number in our data at any given position.
- However, we are more interested in copy numbers changing away from normal and back again.
- Therefore, CBS applies this test:
  - $Z_{ij} = \left(\frac{1}{j-i} + \frac{1}{n-j+i}\right)^{-1/2} \times \left(\frac{S_j - S_i}{j-i} - \frac{S_n - S_j + S_i}{n-j+i}\right)$
    - Or in other words: is there a difference in copy number (log ratios) between the interval enclosed by i-j and the rest of the genome?
    - The first term is maximal for an even partition of the genome (ie the statistic favours large changes over single bin aberrations).
- CBS finds the most significant such change and then is applied recursively to the two segments (the 'rest of the genome' is one segment as CBS circularises all DNA).

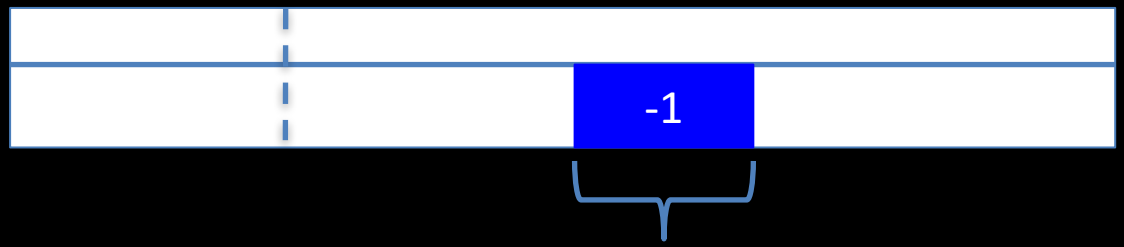
CBS Example

$$Z_{ij} = \left(\frac{1}{j-i} + \frac{1}{n-j+i}\right)^{-1/2} \times \left(\frac{S_j - S_i}{j-i} - \frac{S_n - S_j + S_i}{n-j+i}\right)$$

Initial sequence of log ratios



Interval I,j with maximum difference in average log ratios: slightly negative outside, distinctly positive inside:
(Sj-Si)/(j-i)=(1+2*0.58)/3=0.72. (Sn-Sj+Si)/(n-j+i)=-1/6=-0.17
Note that the formula prefers longer intervals over shorter, so the -1 block has a smaller first term that outweighs the larger second (-1).



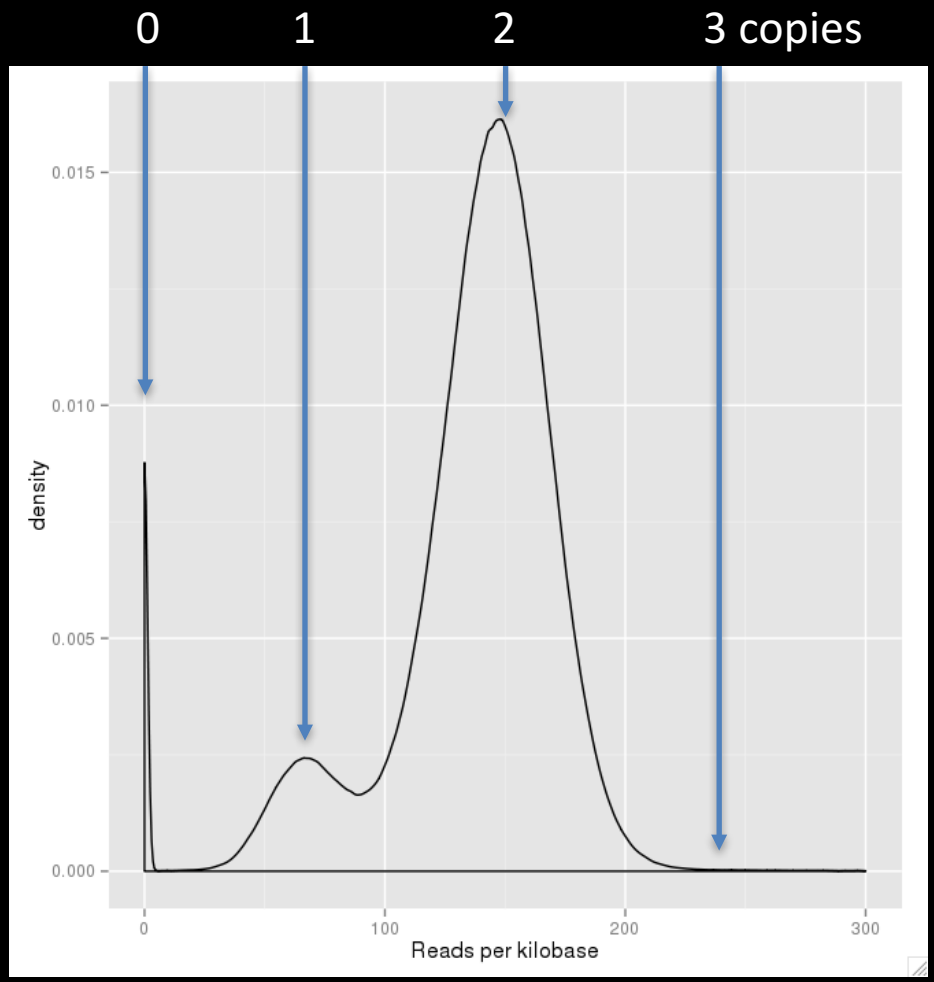
The remaining block gets circularized and the copy number loss gets identified via the Z statistic

CNV Recap

- CNVs are a sub-category of SVs – that change the copy number of alleles.
  - Duplications and deletions.
- Unlike SNVs, we are not interested, where reads are different from the reference.
  - We try to identify areas that have more or fewer reads aligning to them than expected.
- Circular binary segmentation (CBS) assumes the same copy number for all segments (bins), and then identifies the longest consecutive stretch of bins with the highest log-fold change.
  - This procedure is applied recursively, until no significant changes are found anymore.

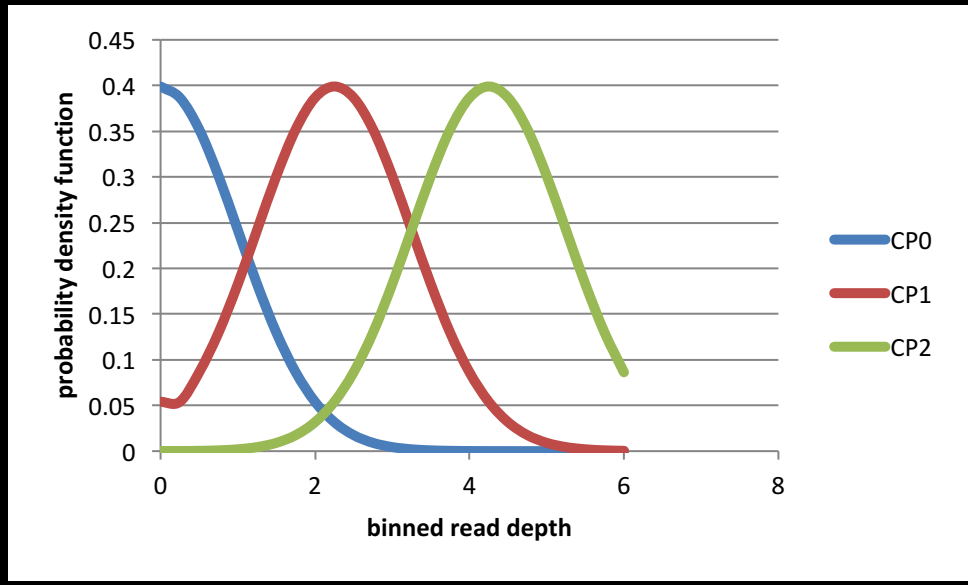
Explicit modelling of CN states

- Instead of investigating "anything that is not normal" (ie copy number 2),
- We can also explicitly model copy number states.



Example Strategy to Call CNVs with HMMs

- Create HMM with hidden states reflecting the copy number of the sample.
- Observations are binned read depths.
- Assume Gaussian distributions for read depth output with means=0, 2, 4 and sd=1



More Modeling Assumptions for HMM

- Assume the following transition probabilities:

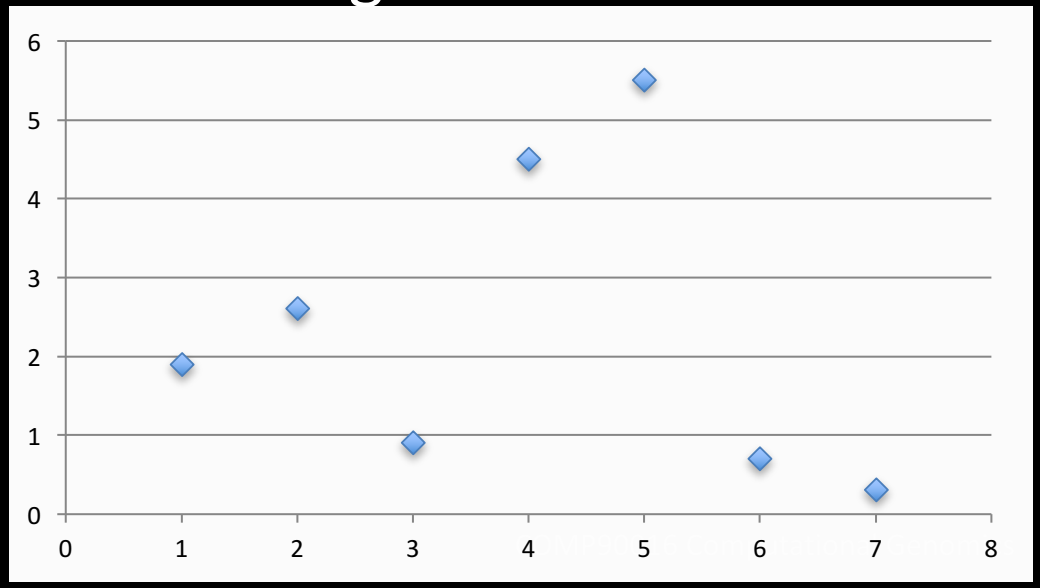
	CP0	CP1	CP2
CP0	0.8	0.1	0.1
CP1	0.1	0.8	0.1
CP2	0.1	0.1	0.8

- Bin the emission probabilities into sections of integer read depths:

	1	2	3	4	5	6
CP0	0.54	0.38	0.07	0.004	0.0002	0.000002
CP1	0.08	0.3	0.37	0.19	0.04	0.002
CP2	0.0005	0.01	0.09	0.3	0.37	0.19

Using the HMM

- For any set of binned read counts:
  - Transform values to our binned emission states (1-6)
  - Calculate most likely path through HMM with Viterbi algorithm.



1	1.9	2
2	2.6	3
3	1.0	2
4	4.5	5
5	5.5	6
6	0.7	1
7	0.3	1

Viterbi Algorithm

$V_{tk} = P(y_t \mid k) * \max x (a_{x,k} * V_{t-1,x})$
a_{xk} = 80% stay, 10%/10% change

	1	2	3	4	5	6
CP0	0.54	0.38	0.07	0.004	0.0002	0.000002
CP1	0.08	0.3	0.37	0.19	0.04	0.002
CP2	0.0005	0.01	0.09	0.3	0.37	0.19

Initialization: 0.33 * P(2|k)

	2	3	2	5	6	1	1
CP0	0.126666667						
CP1	0.1						
CP2	0.003333333						

Viterbi Algorithm

$V_{tk} = P(y_t \mid k) * \max x (a_{x,k} * V_{t-1,x})$
a_{xk} = 80% stay, 10%/10% change

	1	2	3	4	5	6
CP0	0.54	0.38	0.07	0.004	0.0002	0.000002
CP1	0.08	0.3	0.37	0.19	0.04	0.002
CP2	0.0005	0.01	0.09	0.3	0.37	0.19

First iteration: most likely path through two states
CP0->CP0 (0.7%), CP1->CP1 (3%), CP1->CP2 (0.1%)

	2	3	2	5	6	1	1
CP0	0.12666666	0.00709333					
CP1	0.1	0.0296					
CP2	0.00333333	0.00114					

Viterbi Algorithm

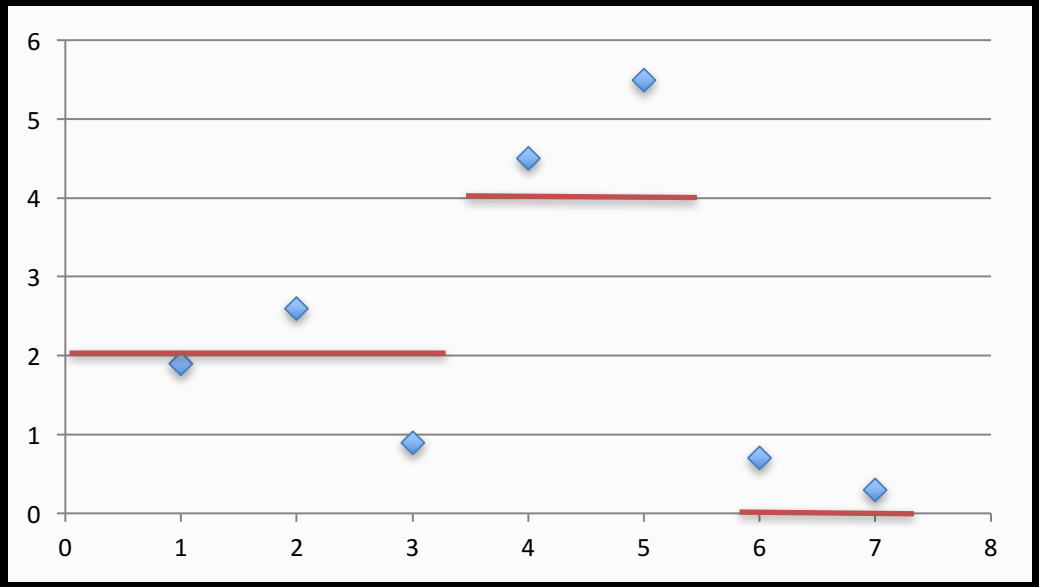
$V_{tk} = P(y_t \mid k) * \max x (a_{x,k} * V_{t-1,x})$
a_{xk} = 80% stay, 10%/10% change

	1	2	3	4	5	6
CP0	0.54	0.38	0.07	0.004	0.0002	0.000002
CP1	0.08	0.3	0.37	0.19	0.04	0.002
CP2	0.0005	0.01	0.09	0.3	0.37	0.19

Final matrix. Most likely path:
CP1->CP1->CP1->CP2->CP2->CP0->CP0

	2	3	1	5	6	1	1
CP0	0.12666666	0.00709333	0.002156373	3.10518E-07	5.25696E-11	2.15746E-06	9.32021E-07
CP1	0.1	0.0296	0.007104	0.000227328	3.63725E-07	3.19623E-07	2.04559E-08
CP2	0.00333333	0.00114	0.0000296	0.000262848	3.99529E-05	1.59812E-08	1.07873E-10

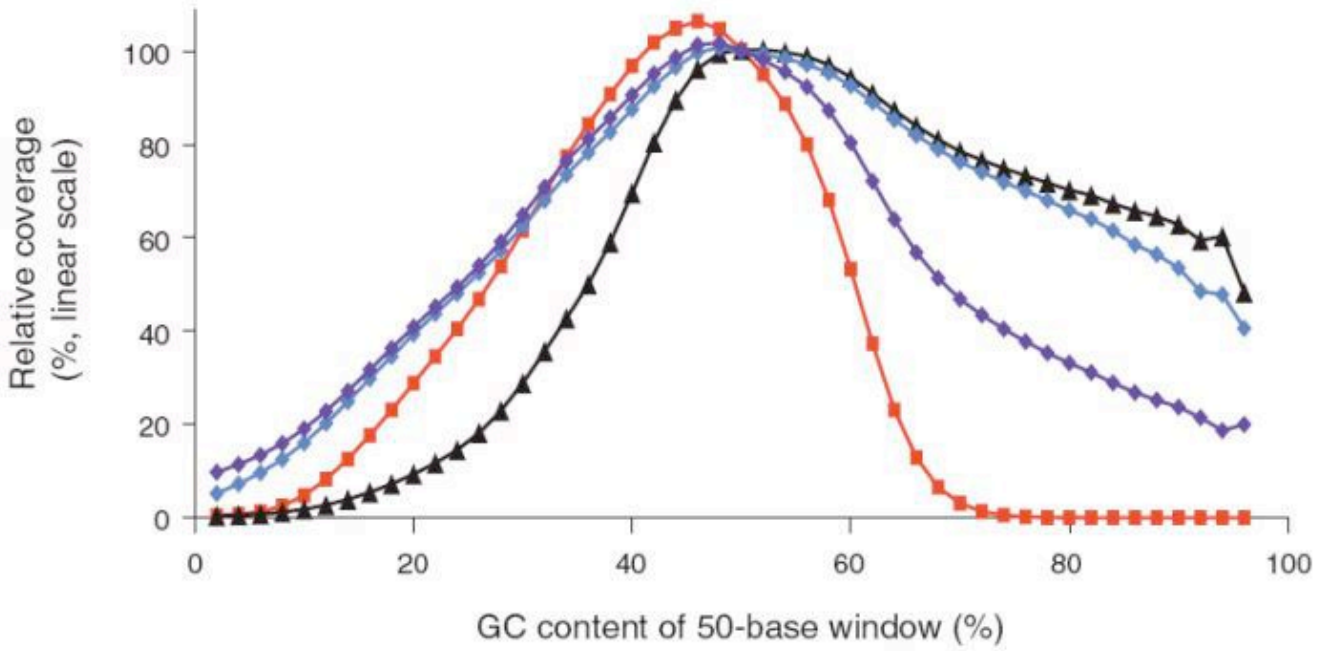
Copy Number Prediction with HMMs



- In real implementations:
  - Log ratios instead of absolute counts to make model scalable.
  - Log probabilities in HMM for easier calculations and to avoid underflow
  - Different copy number states.

Problems that Affect the Read Depth Signal

- GC-content: Areas of very high or very low GC content are difficult to sequence.



Red = standard PCR protocol
Other colors = modified PCR protocols

From Aird et al., Genome Biology *(2011)*

Problems that Affect the Read Depth Signal

- Mappability: Some areas of a genome are easier to map to than others.

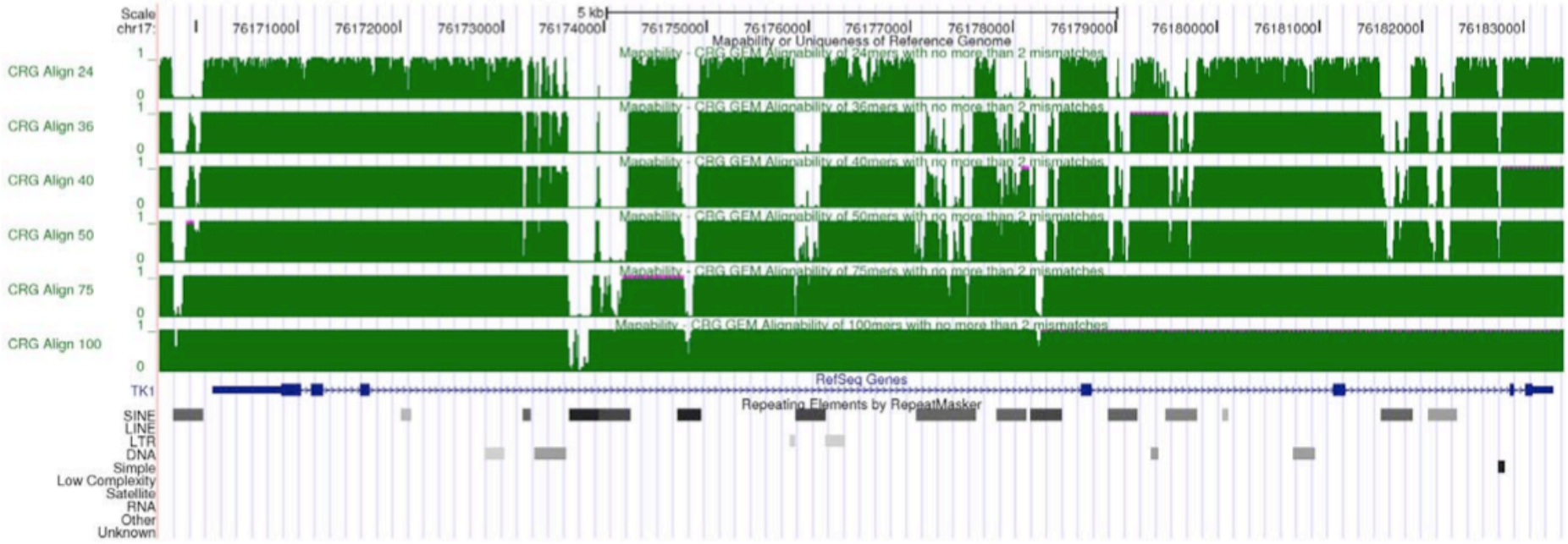


Image from: http://wiki.bits.vib.be

Possible Solutions to non-uniform Read Depth

- Normalisation:
  - Scale the signal according to the GC content or mappability.
- Pre-filtering:
  - Exclude areas of low (or no) mappability from the analysis of the signal.
- Matched normal comparison:
  - If we have DNA from the same genome (patient) that should be basically free of CNVs, we can look at relative bin weights instead of absolute (log ratio over bins, instead of bins and a mean).

Example of CNV detection with sequencing read depth

Read depth signal of 50k bins of low-coverage whole genome sequencing data – a tumour sample

ControlFreec calls 303 CNVs. For example:

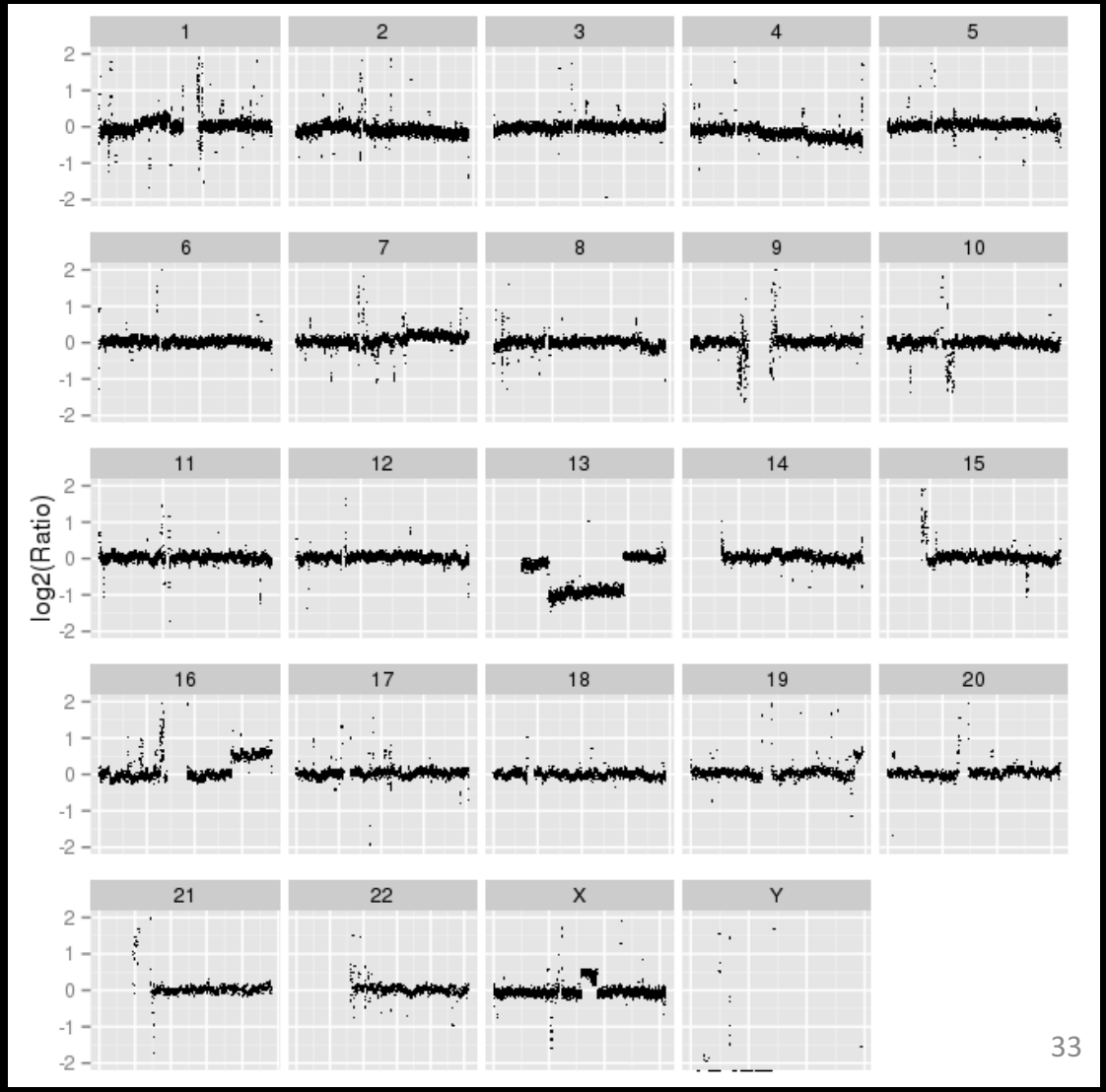
13	36650000	63600000	1	loss
13	63600000	63650000	4	gain
13	63650000	87300000	1	loss

Or

16	69450000	90354753	3	gain

And many small gains.

No events on chr4.

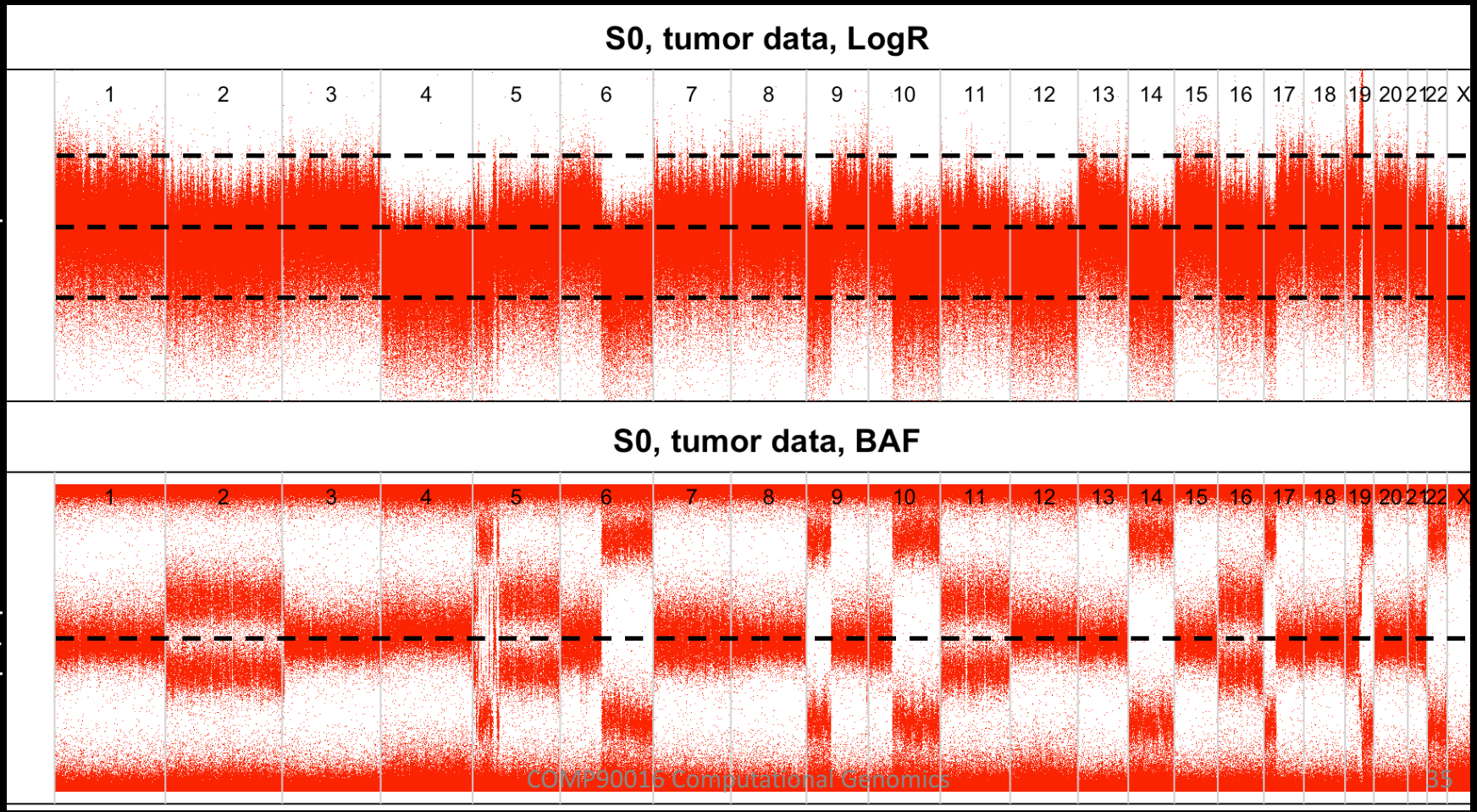


Read Depth SV Detection -- Summary

- A method that somewhat mimics the logR track from the old SNP arrays.
- Has better resolution than arrays, but still pretty coarse: typically in the kilobase range.
- Works for duplications and deletions of genetic material.
  - Answers the question: which areas are deleted or duplicated?
  - But not: If a region is duplicated, where in the genome is it situated?

SNP Arrays & CNVs

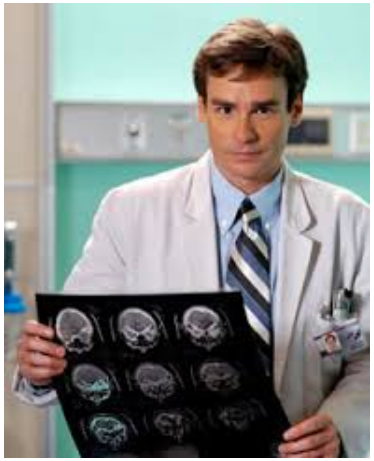
- Tumour SNP arrays data and CNVs
- Reminder: LogR:=$\log_2$(I1+I2/mean(I)), BAF:=I1/(I1+I2)

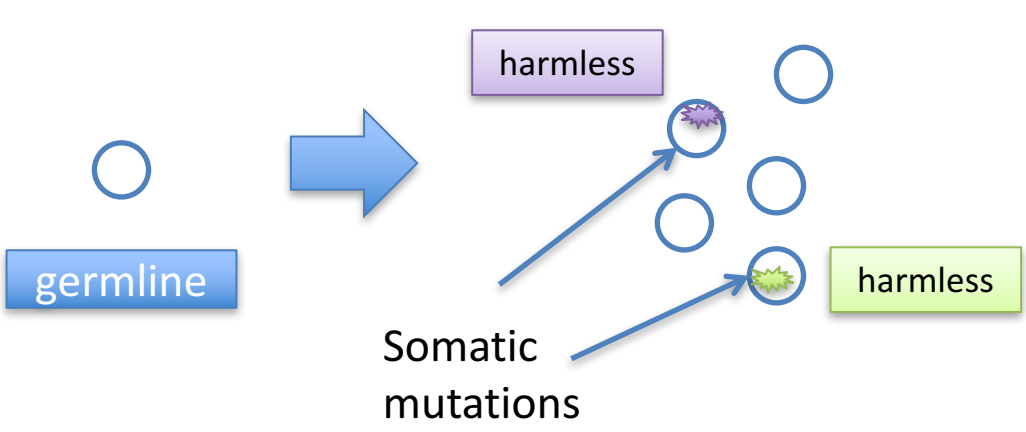


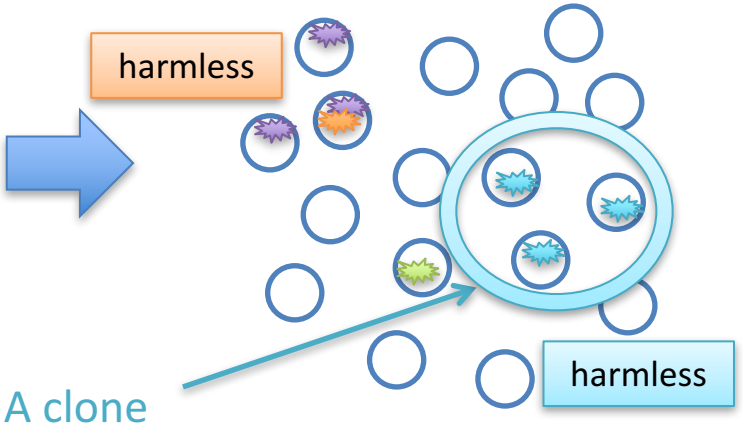
CNVs in a Population of Cells

- There are no partial gains and losses of chromosomes or parts thereof (DNA is either present or absent).
- So, what does a CN of 2.5 mean?
  - It could be that half the population of cells analysed gained a full copy, the other half has 2.
  - Alternatively, a quarter of the cells could have gained 2 copies, and three quarters have 2.
  - Etc.!
- The B allele frequency (BAF) might disambiguate the possible scenarios.

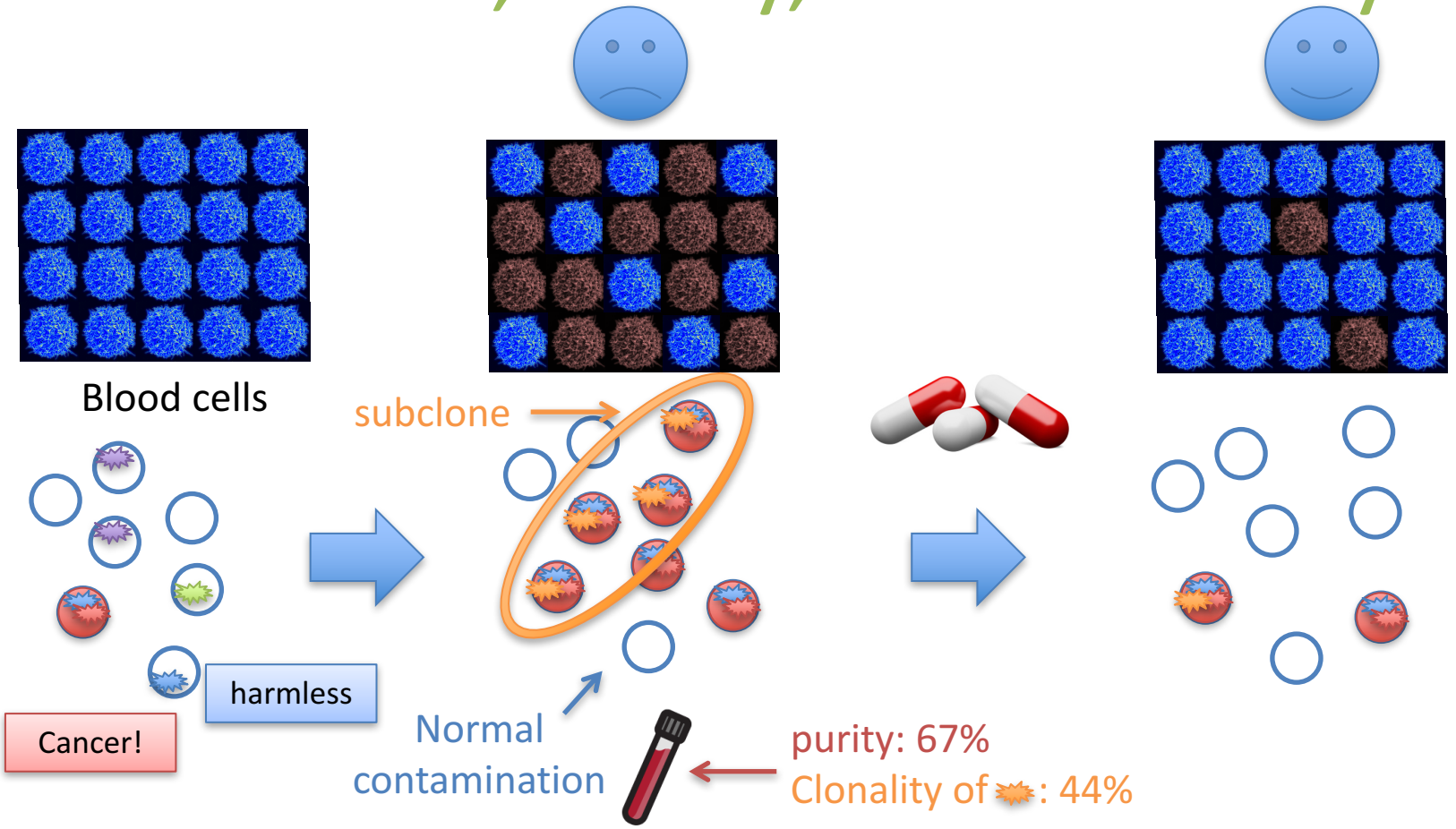
Cancer and Clonality





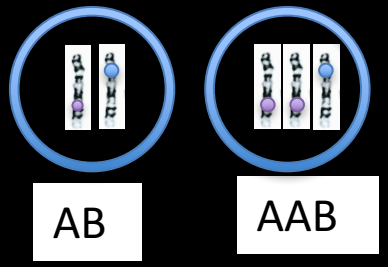


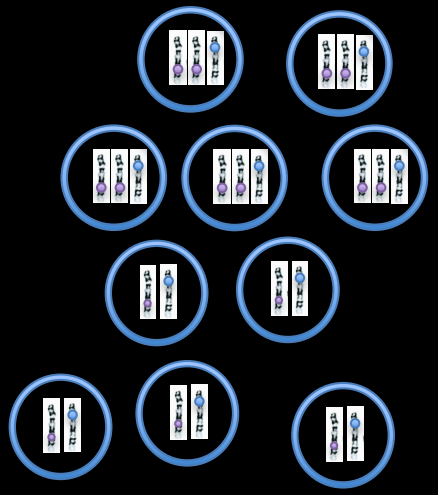
Cancer, Purity, and Clonality



CNVs and Clonality

- Heterozygous germline SNP frequencies affected.
- Gained chromosome 7 (AAB) in all cells:
  - 1.5x coverage (overall 3 copies)
  - SNPs: 50% → 33% or 67%
- Gained chr 7 (AAB) at 50% clonality:
  - 1.25x coverage (overall 2.5 copies)
  - SNPs: 50% → 40% or 60%





Clonality and logR/BAF

- LogR:=$\log_2$((I1+I2)/mean(I)), BAF:=I1/(I1+I2)
- In the context of clones and purity – let us give each allele a copy number and a common clonality.
  - Let n be the average intensity of a regular (1 copy) allele.
  - Let c be the clonality of the following (potentially) variant alleles: vA the copy number of the paternal allele and vB the CN of the maternal allele.
  - It follows for a diploid chromosome:

$$\log R = \log 2\left(\frac{(vA+vB)cn+(1-c)2n}{2n}\right) = \log 2\left(\frac{((vA+vB)c+2-2c)}{2}\right)$$

$$BAF = \frac{vB\cdot c\cdot n+(1-c)n}{(vA+vB)cn+(1-c)2n} = \frac{vB\cdot c+1-c}{(vA+vB)c+2-2c}$$

  - Note, the BAF is written down for *heterozygous* variants only! Homozygous variants *ALWAYS* have frequency 0 or 1, no matter what the copy number.
  - This formulation is applicable to SNP arrays and sequencing data alike.

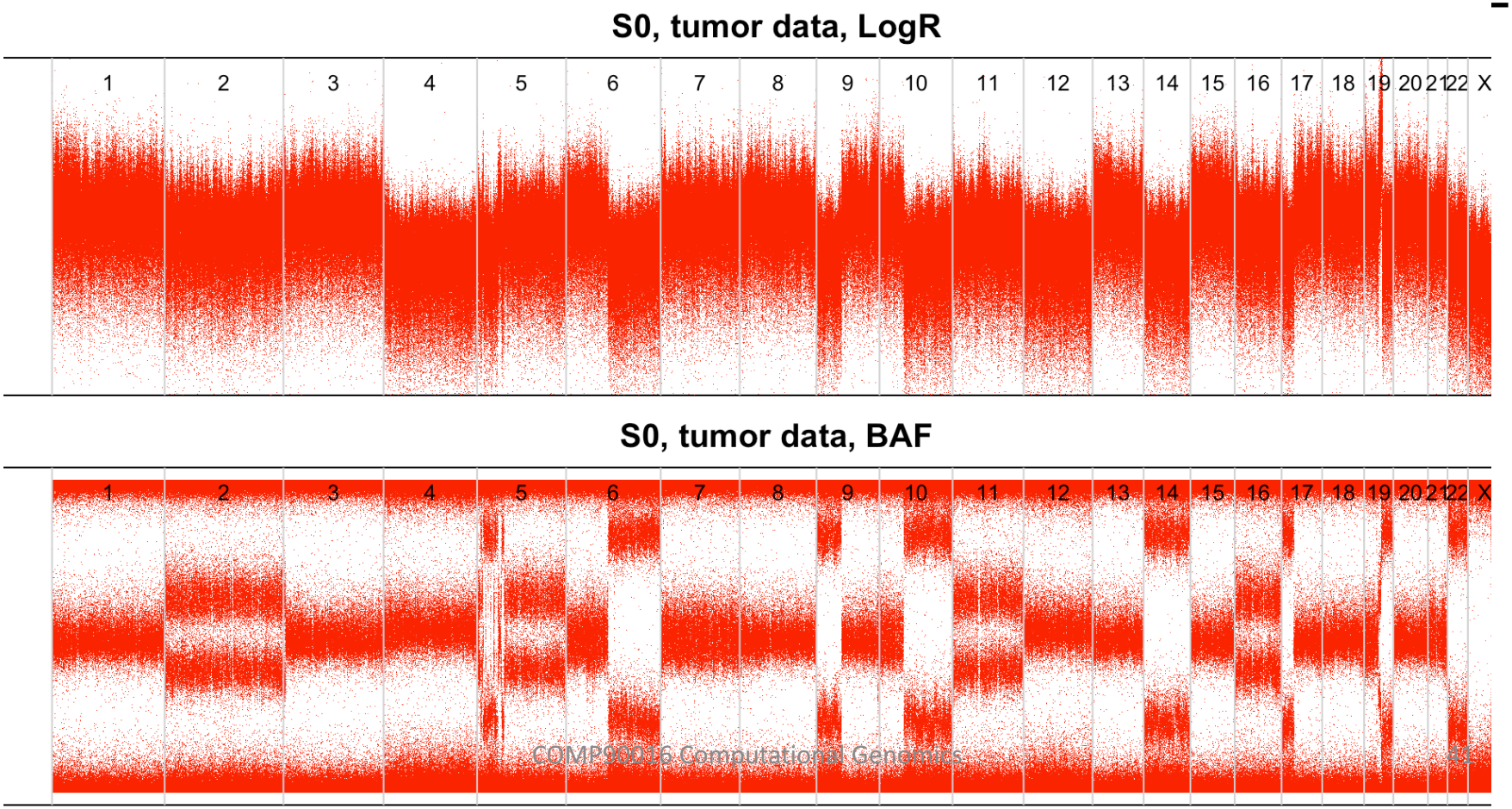
s. In

vA and vB are integers (full copies of chromosomes)!
0<=c<=1 – the ratio of the cells containing the clone.

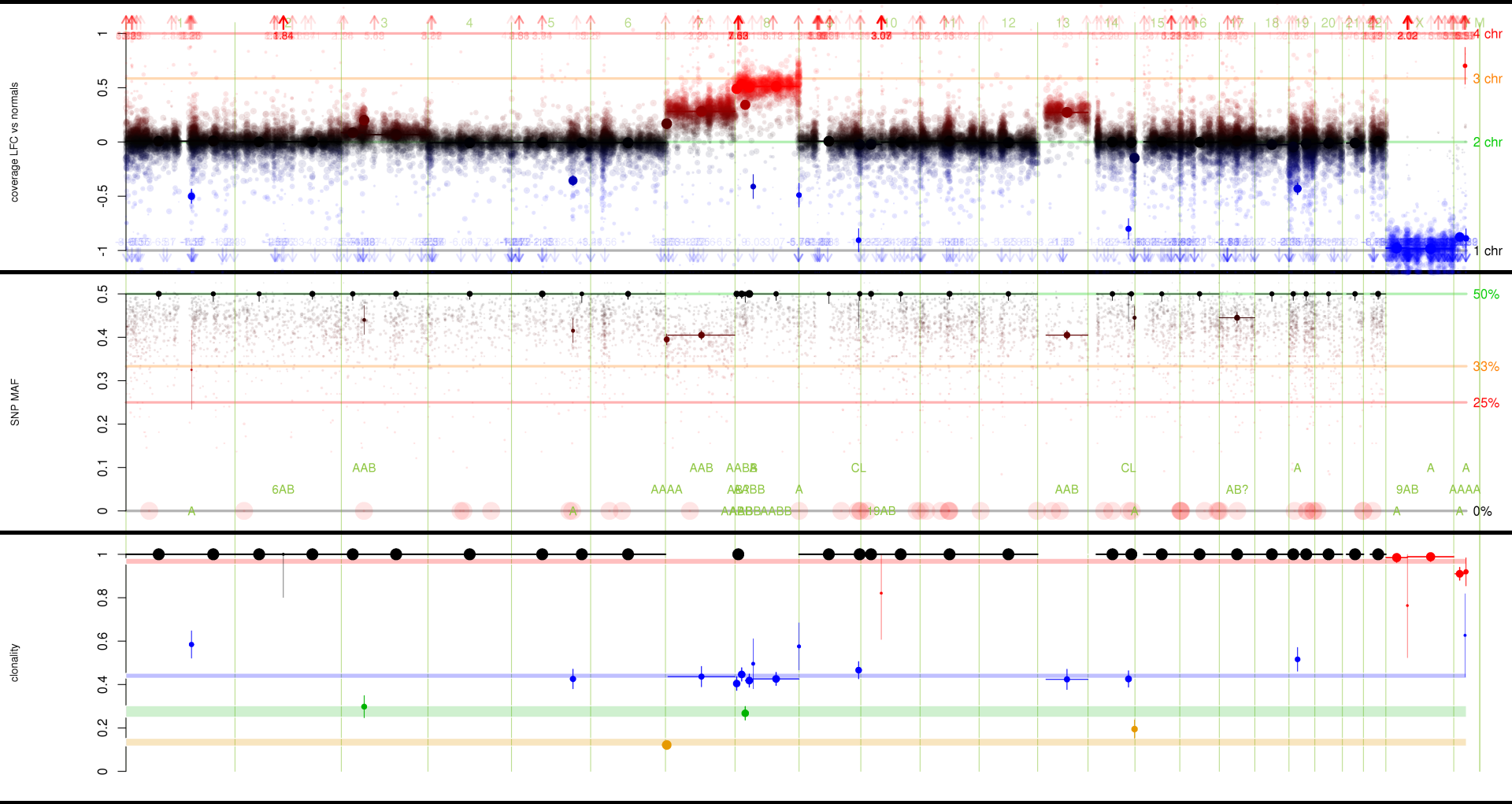
Example: CNV detection with SNP arrays (the sub-clonality issue)

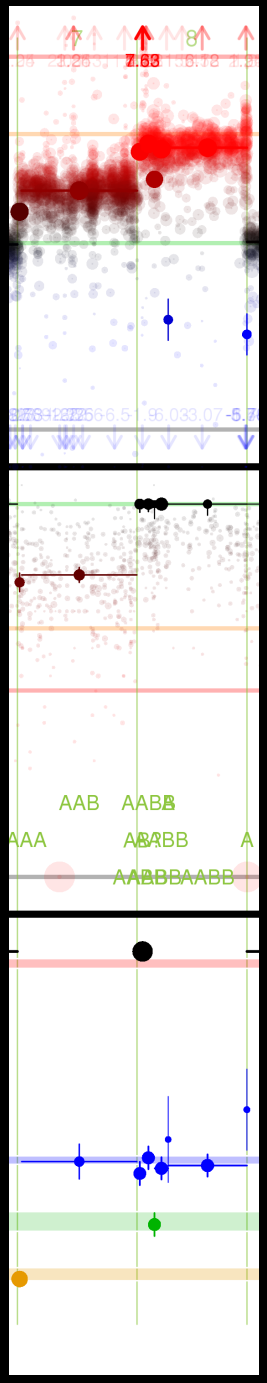
¼ cells lost a copy of chr2 -> splitting of the 50% band

C1	C2	C3	C4	BAF
AA	AA	AA	A	0%
BB	BB	BB	B	100%
AB	AB	AB	A	43%
AB	AB	AB	B	57%



CNVs, Clonality, and Sequencing Data





CNVs and Clones (continued)

- Examine chr7:
  - A non integer gain of DNA is observed across the entire chromosome.
  - > only a sub-population has gained a single copy (or more).
- logR ~ 0.32
  Setting vA=1 and vB=2, we can solve for c:

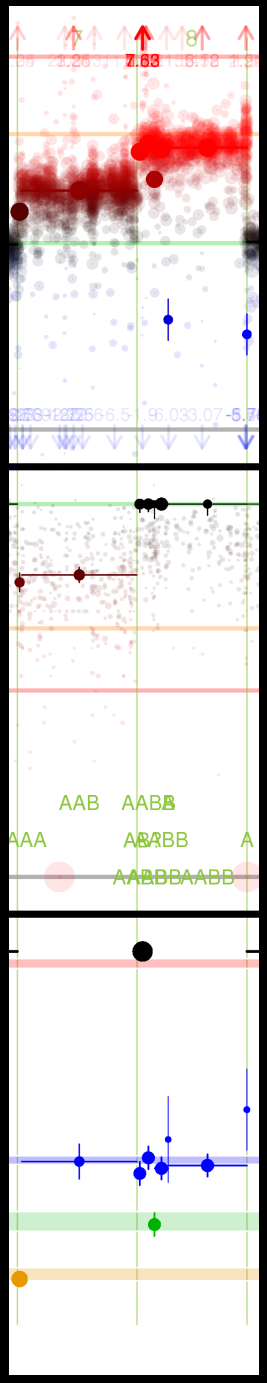
$$0.32 = log_2\left(\frac{(vA + vB)c + 2 - 2c}{2}\right)$$

$$0.32 = log_2\left(\frac{3c + 2 - 2c}{2}\right)$$

$$0.32 = log_2\left(\frac{c + 2}{2}\right)$$

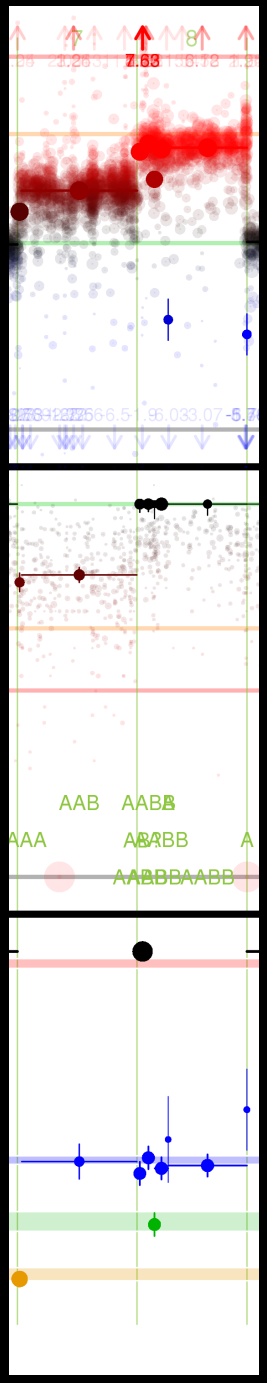
$$2^{0.32} = \left(\frac{c + 2}{2}\right)$$

$$c = 2^{1.32} - 2 = 0.5$$



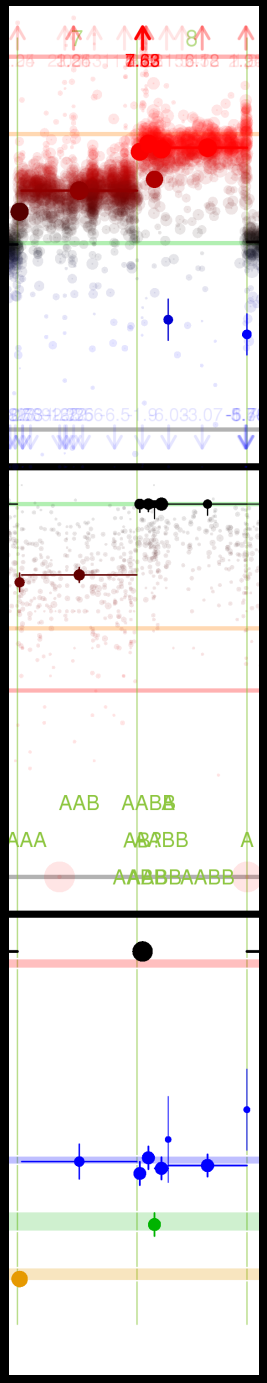
CNVs and Clones (continued)

- Examine chr7:
  - A non integer gain of DNA is observed across the entire chromosome.
  - > only a sub-population has gained a single copy (or more).
- logR ~ 0.32 -> c=0.5. Use this for the BAF formula:
  BAF=60%
- This coincides with the observation of BAF~50+-10%



CNVs and Clones (continued)

- Examine chr8:
  - A full gain of chr8 can be observed.
  - Has the entire population acquired an extra copy?
- logR ~ 0.58
  Setting vA=1 and vB=2, we get c=1.
- It follows BAF = 66.66%
  However, we do not observe this!
  We observe BAF = 50%.



CNVs and Clones (continued)

- Examine chr8:
  - A full gain of chr8 can be observed.
  - Has the entire population acquired an extra copy?
- logR ~ 0.58
  Setting vA=2 and vB=2, we get

$$0.58 = log_2\left(\frac{(vA+vB)c+2-2c}{2}\right)$$

$$0.58 = log_2\left(\frac{4c+2-2c}{2}\right)$$

$$0.58 = log_2\left(\frac{2c+2}{2}\right)$$

$$2^{0.58} = \left(\frac{2c+2}{2}\right)$$

$$c = \frac{2^{1.58}-2}{2} = 0.5$$

- And BAF = 50%.
- A solution that is consistent with our observation.

CNV Summary

- CNVs are easiest to detect among the structural variants:
  - The signal only depends on reads mapping to the reference or not, so details such as base quality, nearby SNPs etc. don't matter.
- Similar to SNP calling, CNVs can be detected with a set number of discrete copy number states in mind.
  - See HMM example.
  - Comparable to genotyping (heterozygous/homozgyous)
- Or on a continuous level:
  - Important in cancer and heterogeneous samples, where "non-integer" changes are observed.
  - Comparable to SNV calling.
  - See CBS.
- The integration of BAF into CNV calling is – surprisingly – rarely done!
  - An open area of research.
- Other signal is needed to investigate where the CN changes take place in the genome.

Fluorescence *in situ* hybridization (FISH)

Stain DNA with fluorescent material according to the chromosome of origin.

Chromosomal rearrangements light up as multi-colored chimeras.