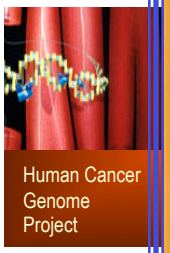
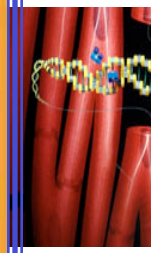
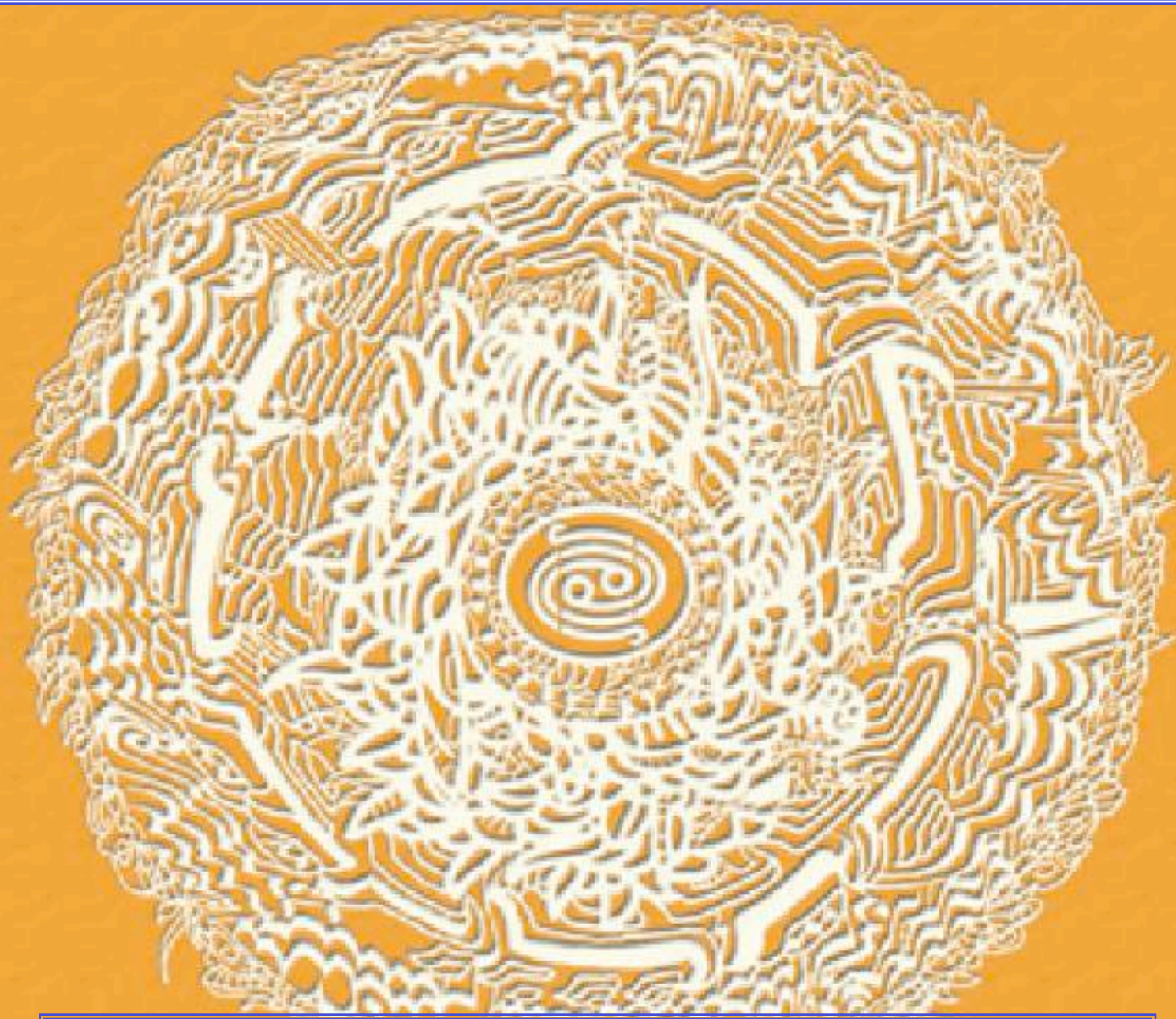
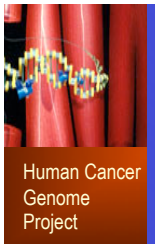


Human Cancer Genome Project

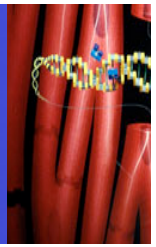




Bioinformatics/Genomics of Cancer:



Bud Mishra



Professor of Computer Science, Mathematics and Cell Biology

Courant Institute, NYU School of Medicine, Tata Institute of Fundamental Research, and Mt. Sinai School of Medicine

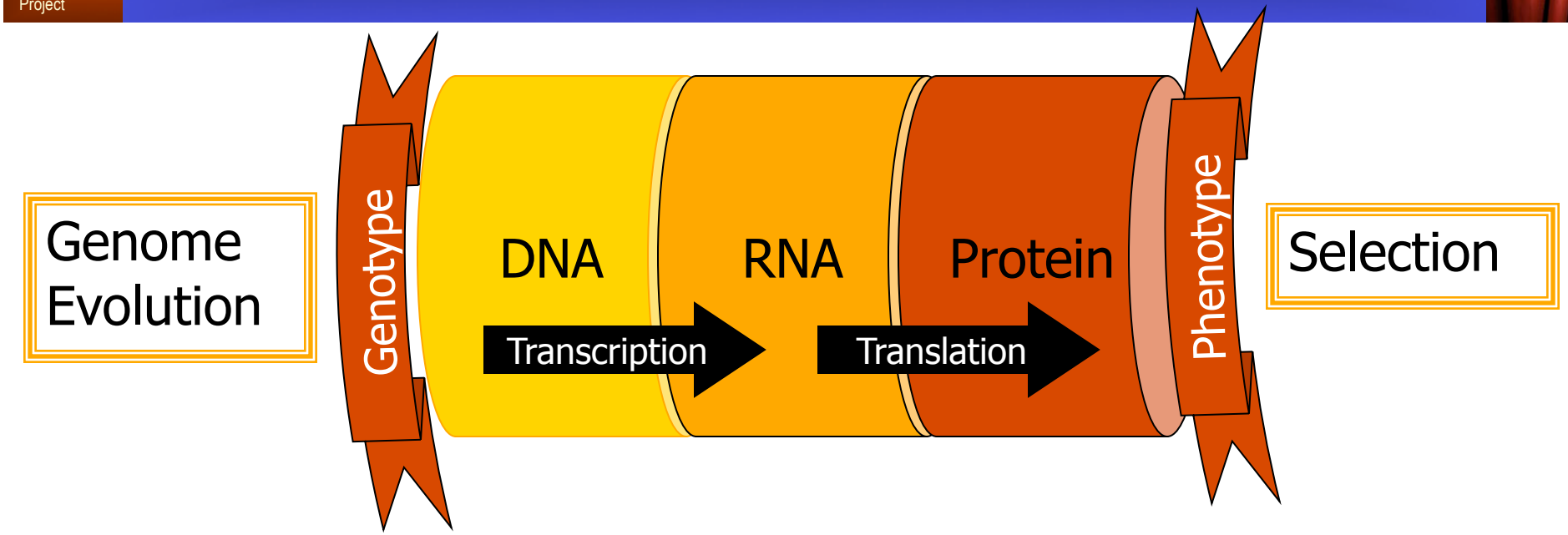
Introduction: Cancer and Genomics:

What we know & what we do not

"Cancer is a disease of the genome."

Human Cancer Genome Project

The New Synthesis



Part-lists, Annotation, Ontologies

Human Cancer Genome Project

Cancer Initiation and Progression

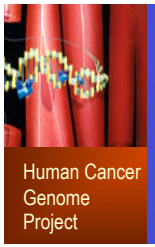
Mutations, Translocations, Amplifications, Deletions

Epigenomics (Hyper & Hypo-Methylation)

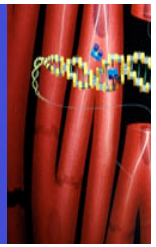
Alternate Splicing

Cancer Initiation and Progression

Proliferation, Motility, Immortality, Metastasis, Signaling



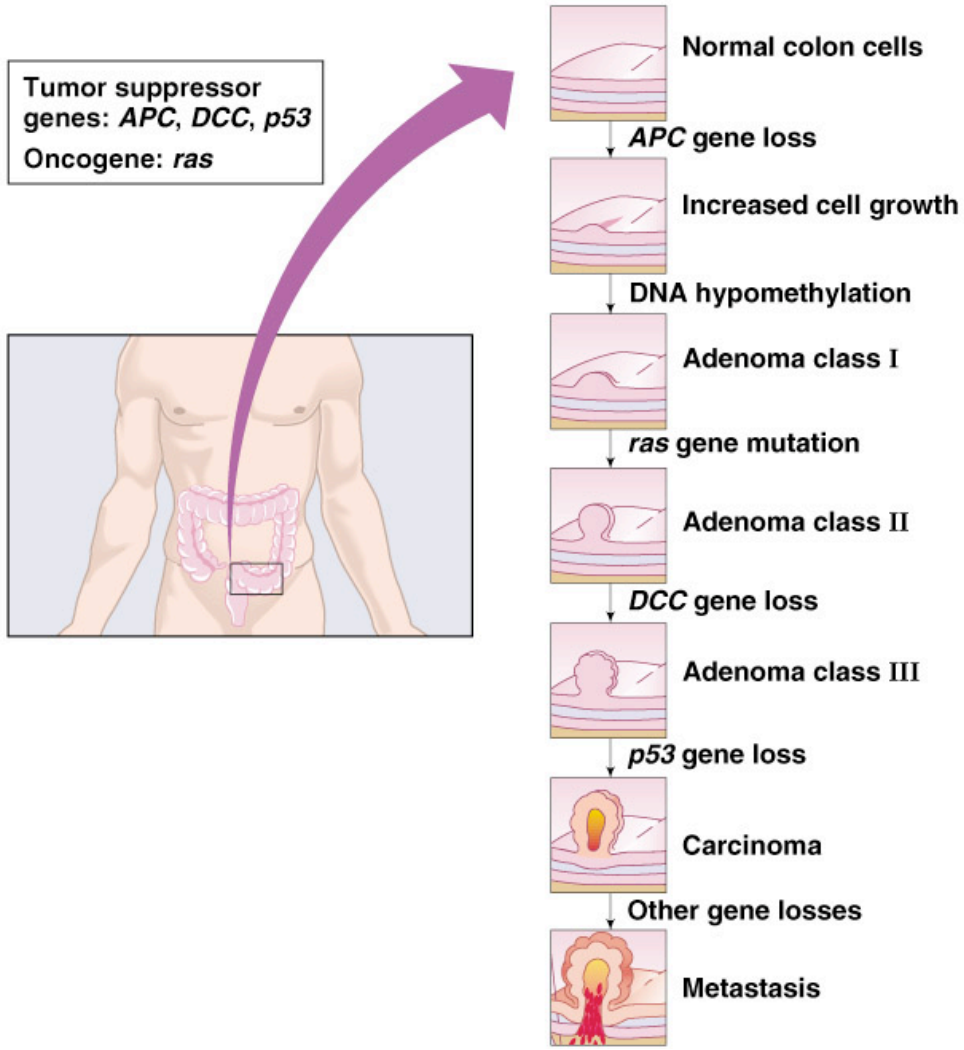
Multi-step Nature of Cancer:



- Cancer is a stepwise process, typically requiring accumulation of mutations in a number of genes.
- ~6-7 independent mutations typically occur over several decades:
  - Conversion of **proto-oncogenes** to oncogenes
  - Inactivation of **tumor suppressor gene**

Amplifications & Deletions

Human Cancer Genome Project



Mutation in a TSG

Epigenomics

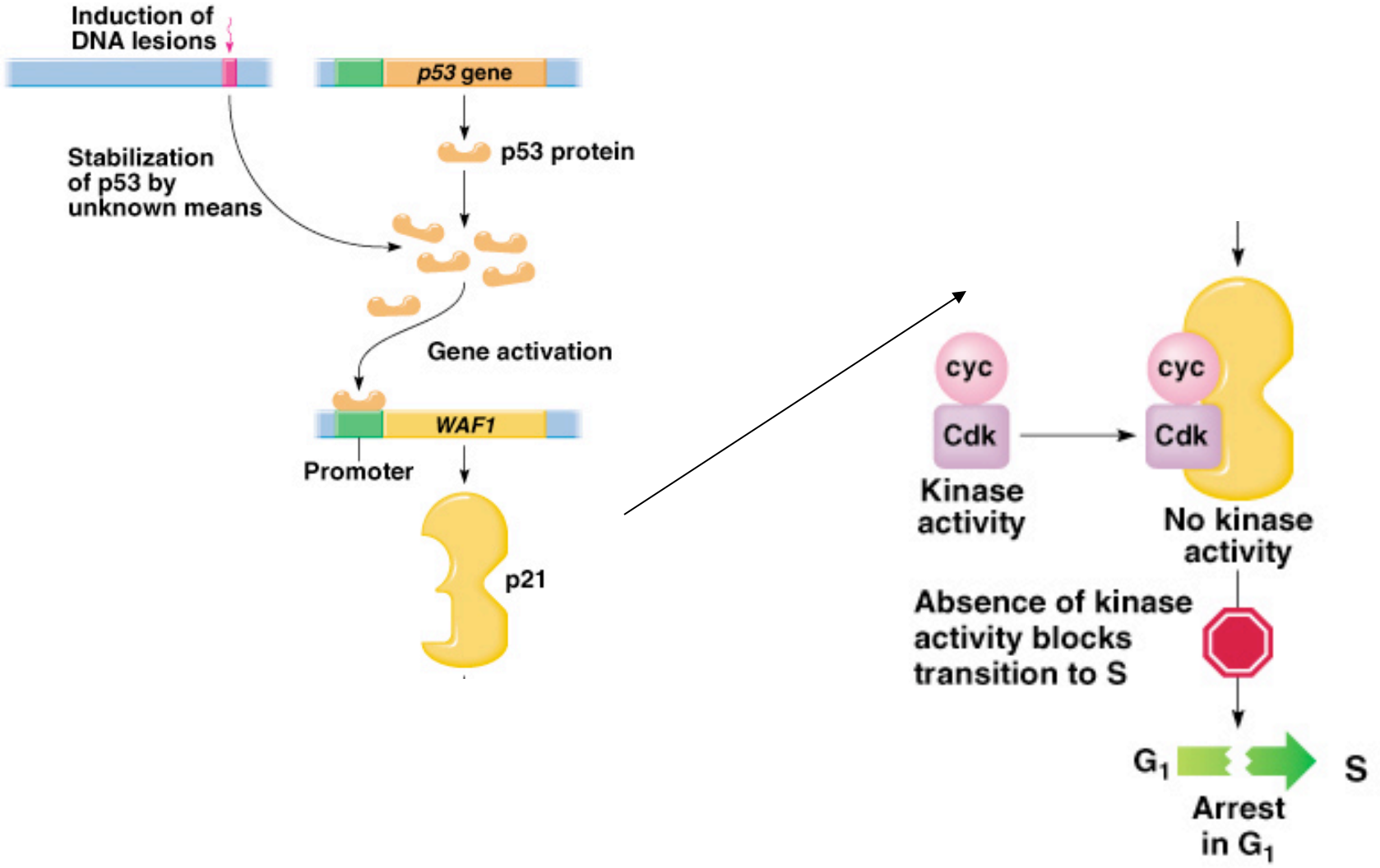
Conversion of a Proto-Oncogene

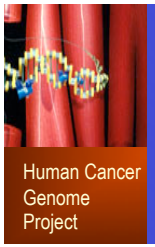
Deletion of a TSG

Deletion of a TSG

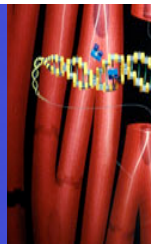
Human Cancer Genome Project

P53 Gene (TSG)

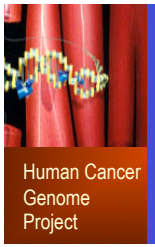




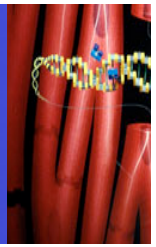
The Cancer Genome Atlas



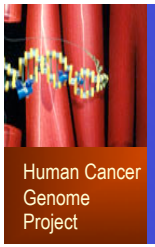
- **Obtain a comprehensive description of the genetic basis of human cancer**.
  - Identify and characterize all the sites of genomic alteration associated at significant frequency with all major types of cancers.



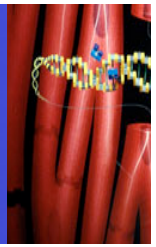
The Cancer Genome Atlas



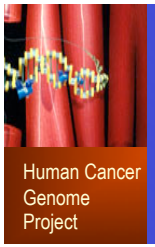
- Increase the effectiveness of research to understand
  - tumor initiation and progression,
  - susceptibility to carcinogensis,
  - development of cancer therapeutics,
  - approaches for early detection of tumors &
  - the design of clinical trials.



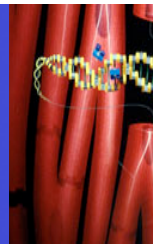
Specific Goals



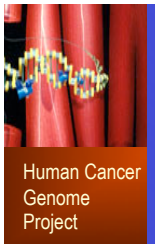
- Identify all genomic alterations significantly associated with all major cancer types.
- Such knowledge will propel work by thousands of investigators in cancer biology, epidemiology, diagnostics and therapeutics.



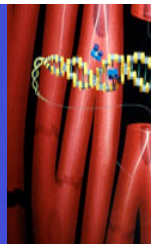
To Achieve this goal …



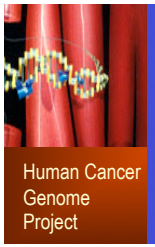
- Create large collection of appropriate, clinically annotated samples from all major types of cancer; and
- Characterize each sample in terms of:
  - All regions of genomic loss or amplification,
  - All mutations in the coding regions of all human genes,
  - All chromosomal rearrangements,
  - All regions of aberrant methylation, and
  - Complete gene expression profile, as well as other appropriate technologies.



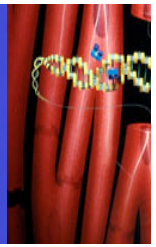
Biomedical Rationale



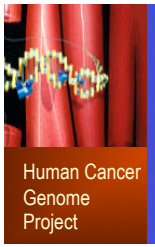
- Cancer is a heterogeneous collection of heterogeneous diseases.
  - For example, prostate cancer can be an indolent disease remaining dormant throughout life or an aggressive disease leading to death.
  - However, we have no clear understanding of why such tumors differ.



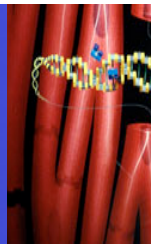
Biomedical Rationale



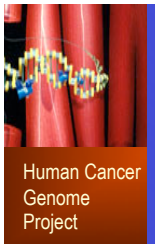
- Cancer is fundamentally a disease of genomic alteration.
  - Cancer cells typically carry many genomic alterations that confer on tumors their distinctive abilities (such as the capacity to proliferate and metastasize, ignoring the normal signals that block cellular growth and migration) and liabilities (such as unique dependence on certain cellular pathways, which potentially render them sensitive to certain treatments that spare normal cells).



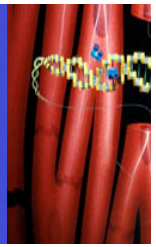
History



- 1960s
  - The genetic basis of cancer was clear from cytogenetic studies that showed consistent translocations associated with specific cancers (notably the so-called Philadelphia chromosome in chronic myelogenous leukemia).
- 1970s
  - Recognize specific cancer-causing mutations through recombinant DNA revolution of the 1970s.
  - The identification of the first vertebrate and human oncogenes and the first tumor suppressor genes,
  - These discoveries have elucidated the cellular pathways governing processes such as cell-cycle progression, cell-death control, signal transduction, cell migration, protein translation, protein degradation and transcription.
- **For no human cancer do we have a comprehensive understanding of the events required.**



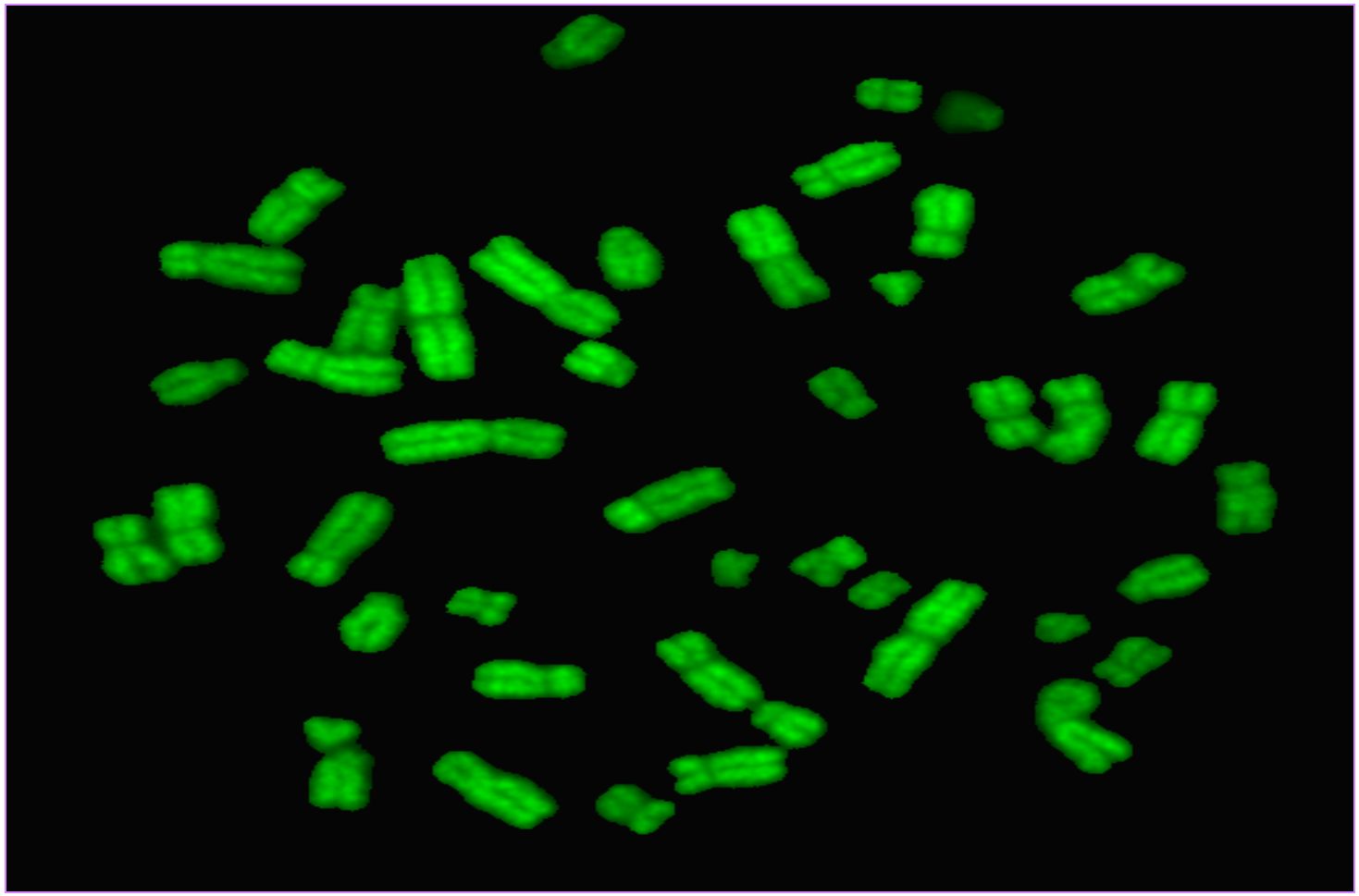
Scientific Foundation for a Human Cancer Genome Project



- **Gene resequencing**.
  - Specific gene classes (such as kinases and phosphatases) in particular cancer types.
- **Epigenetic changes**.
  - Loss of function of tumor suppressor genes by epigenetic modification of the genome — such as DNA methylation and histone modification.
- **Genomic loss and amplification**.
  - Consistent association with genomic loss or amplification in many specific regions, indicating that these regions harbor key cancer associated genes
- **Chromosome rearrangements**.
  - Activate kinase pathways through fusion proteins or inactivating differentiation programs through gene disruption.
  - Hematological malignancies: a single stereotypical translocation in some diseases (such as CML) and as many as 20 important translocations in others (such as AML).
  - Adult solid tumors have not been as well characterized, in part owing to technical hurdles.

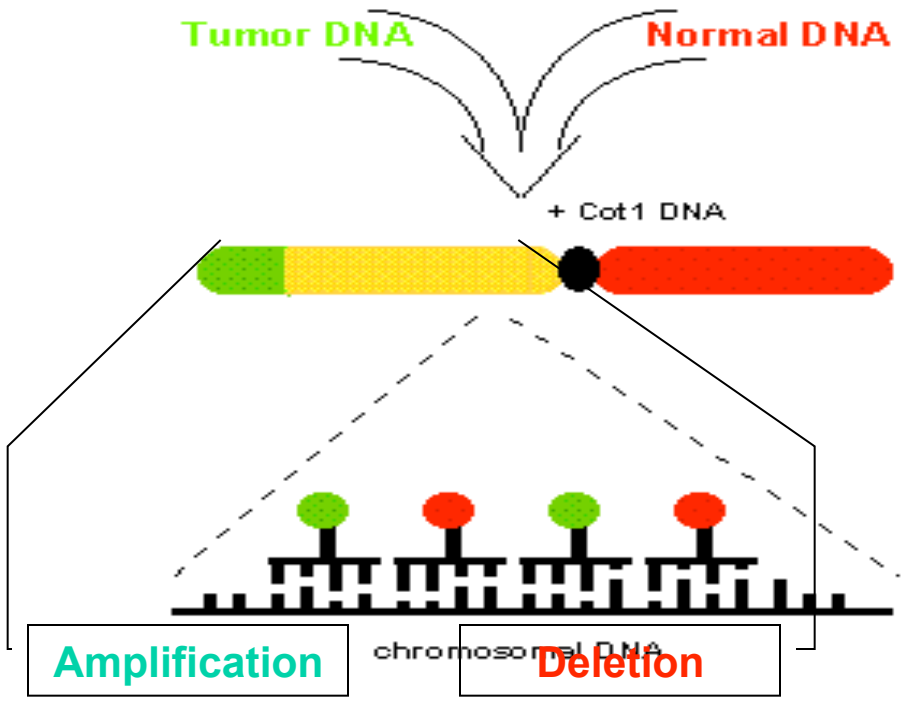
Human Cancer Genome Project

Karyotyping



Human Cancer Genome Project

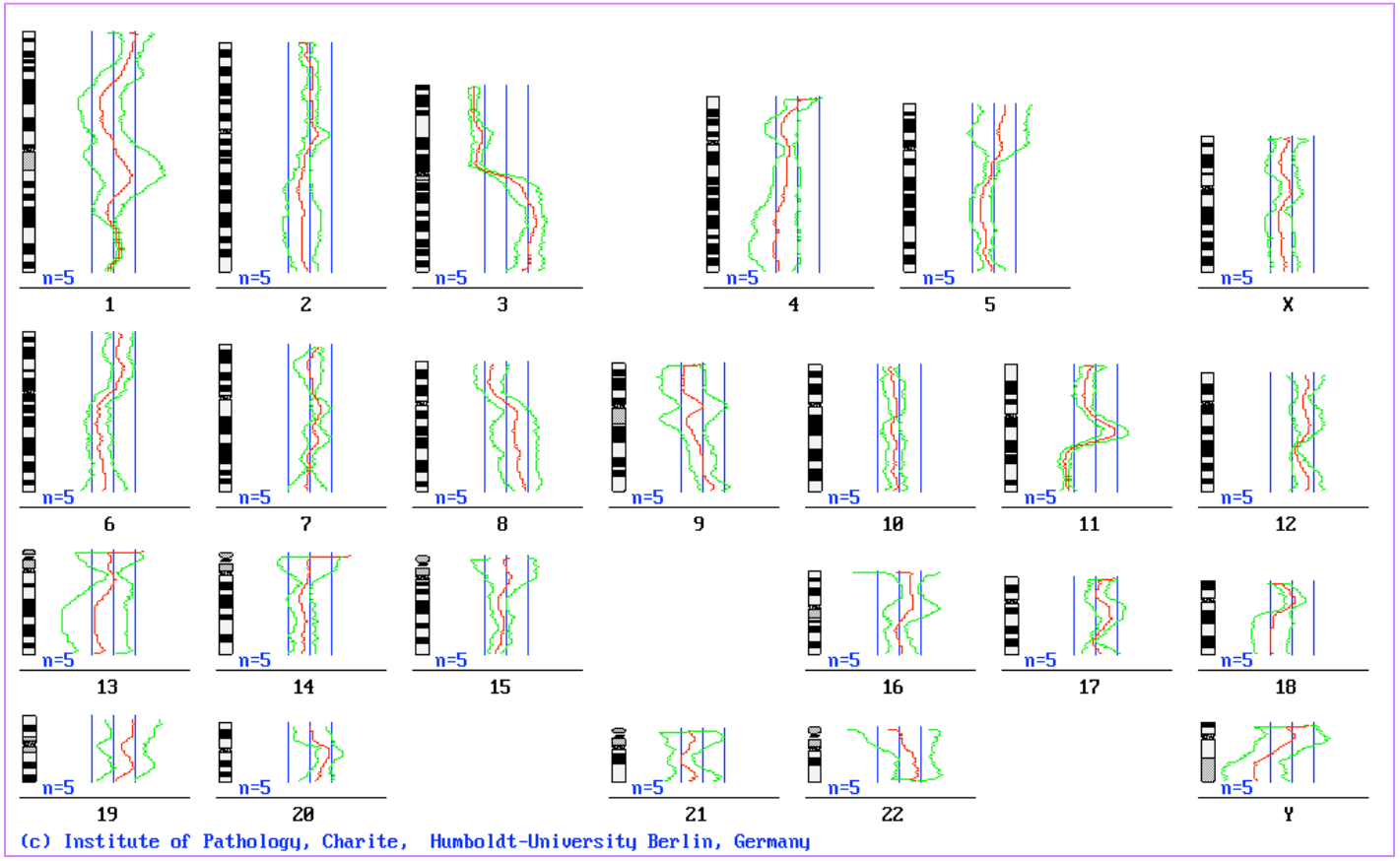
CGH:
Comparative Genomic Hybridization.



- Equal amounts of biotin-labeled tumor DNA and digoxigenin-labeled normal reference DNA are hybridized to normal metaphase chromosomes
- The tumor DNA is visualized with fluorescein and the normal DNA with rhodamine
- The signal intensities of the different fluorochromes are quantitated along the single chromosomes
- The over-and underrepresented DNA segments are quantified by computation of tumor/normal ratio images and average ratio profiles

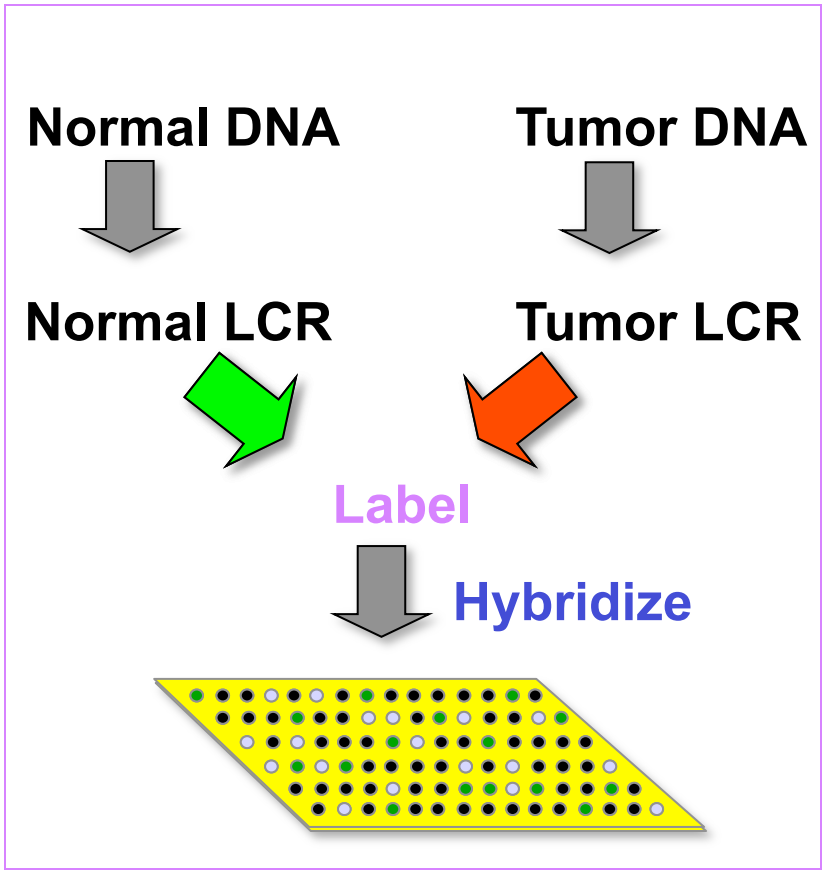
Human Cancer Genome Project

CGH: Comparative Genomic Hybridization.



Microarray Analysis of Cancer Genome

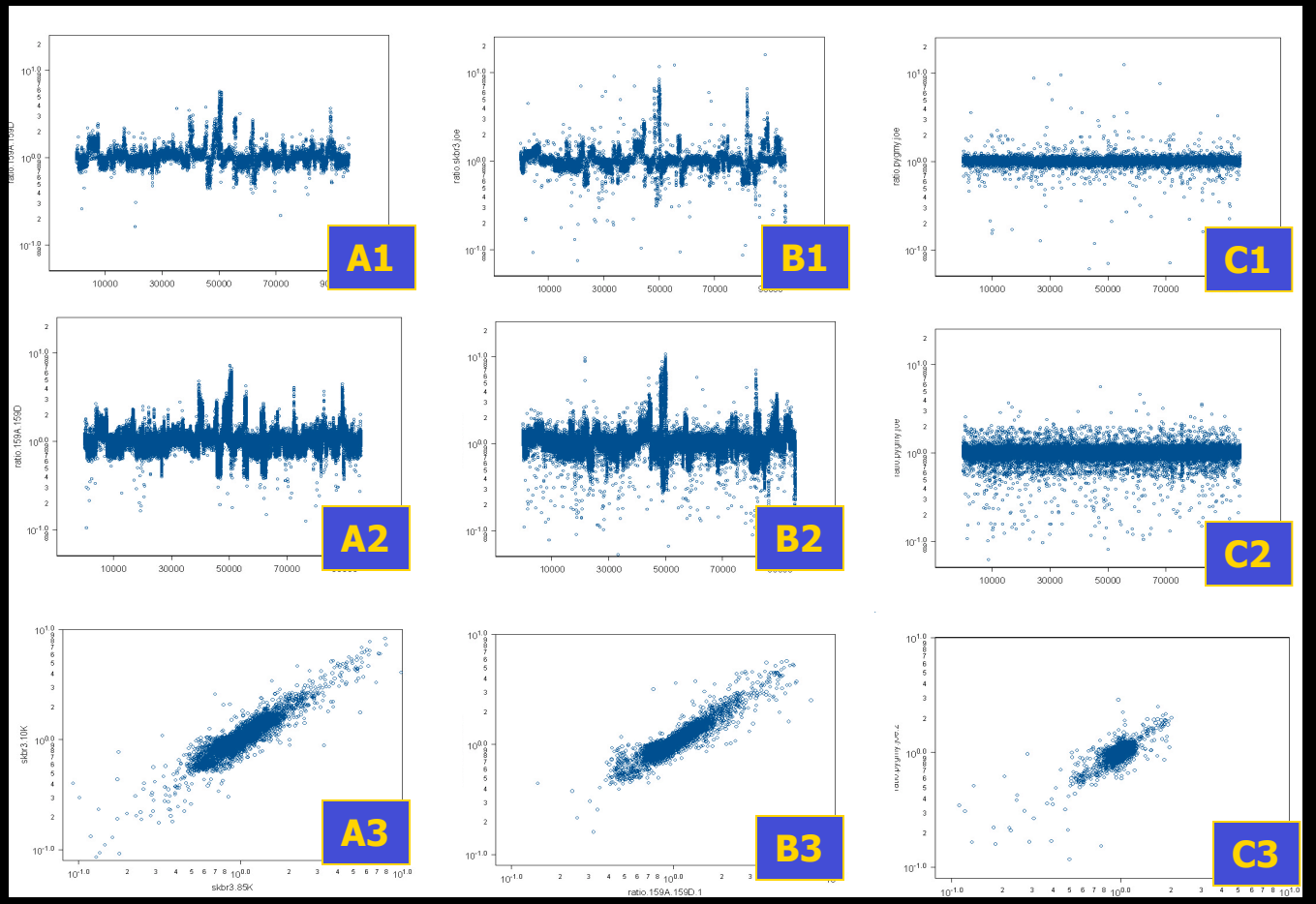
Human Cancer Genome Project



- Representations are reproducible samplings of DNA populations in which the resulting DNA has a new format and reduced complexity.
  - We array probes derived from low complexity representations of the normal genome
  - We measure differences in gene copy number between samples ratiometrically
  - Since representations have a lower nucleotide complexity than total genomic DNA, we obtain a stronger specific hybridization signal relative to non-specific and noise

Human Cancer Genome Project

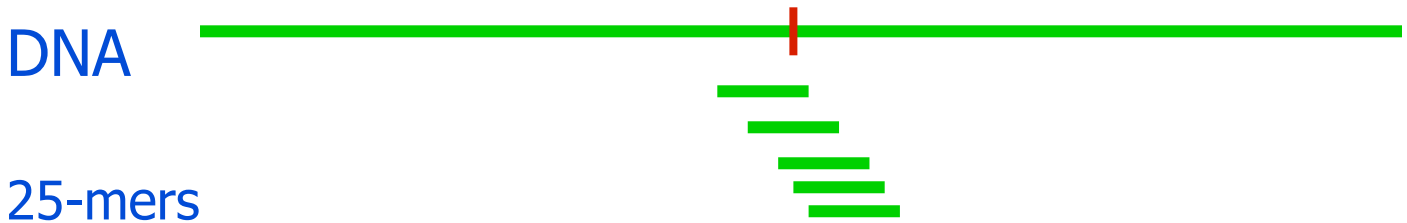
Copy Number Fluctuation



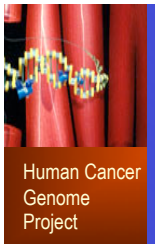
Human Cancer Genome Project

Oligo Arrays: SNP genotyping

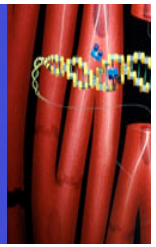
- Given 500K human SNPs to be measured, select 10 25-mers that over lap each SNP location for Allele A.



  - Select another 10 25-mers corresponding to SNP Allele B.
  - Problem : Cross Hybridization



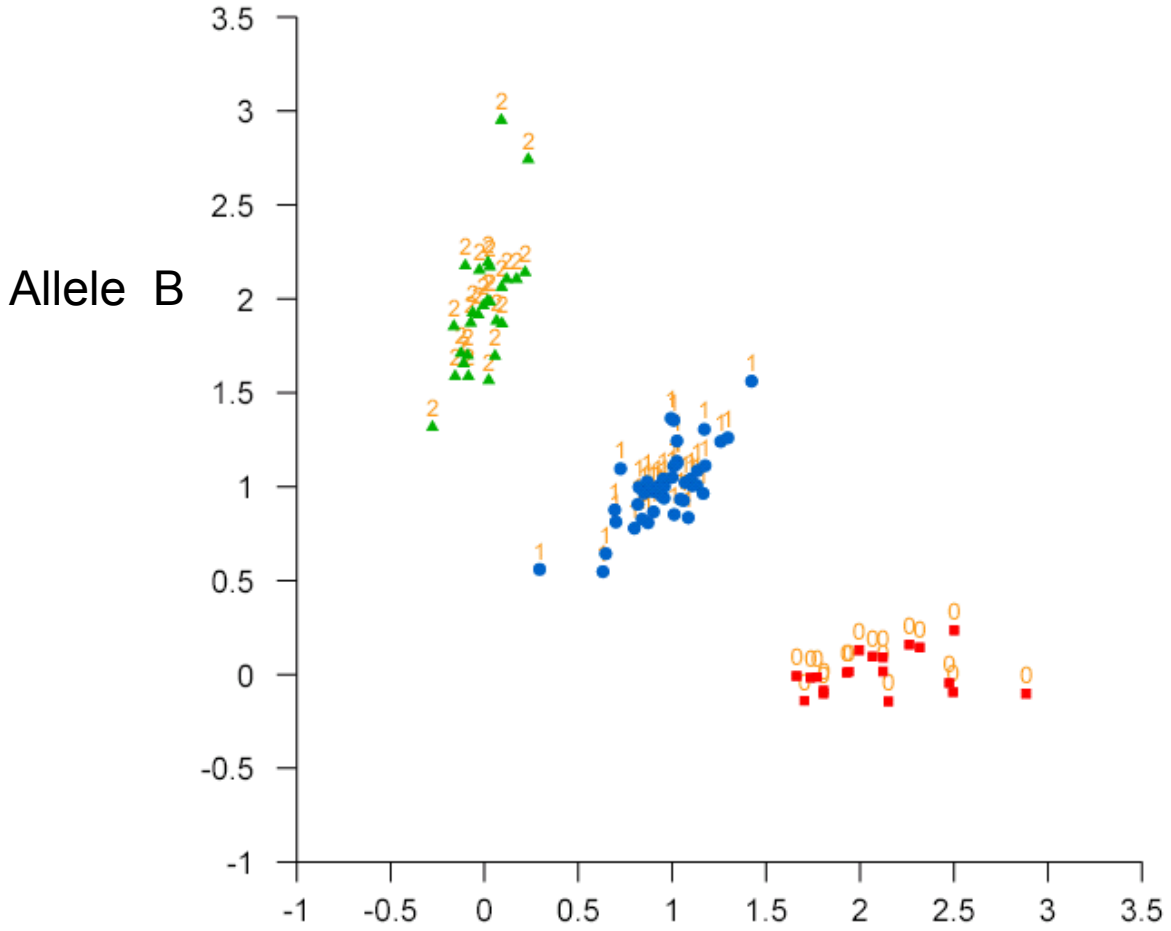
Using SNP arrays to detect Genomic Aberrations



- Each SNP "probeset" measures absense/presence of one of two Alleles.
- If a region of DNA is deleted by cancer, one or both alleles will be missing!
- If a region of DNA is duplicated/ amplified by cancer, one or both alleles will be amplified.
- Problem : Oligo arrays are noisy.

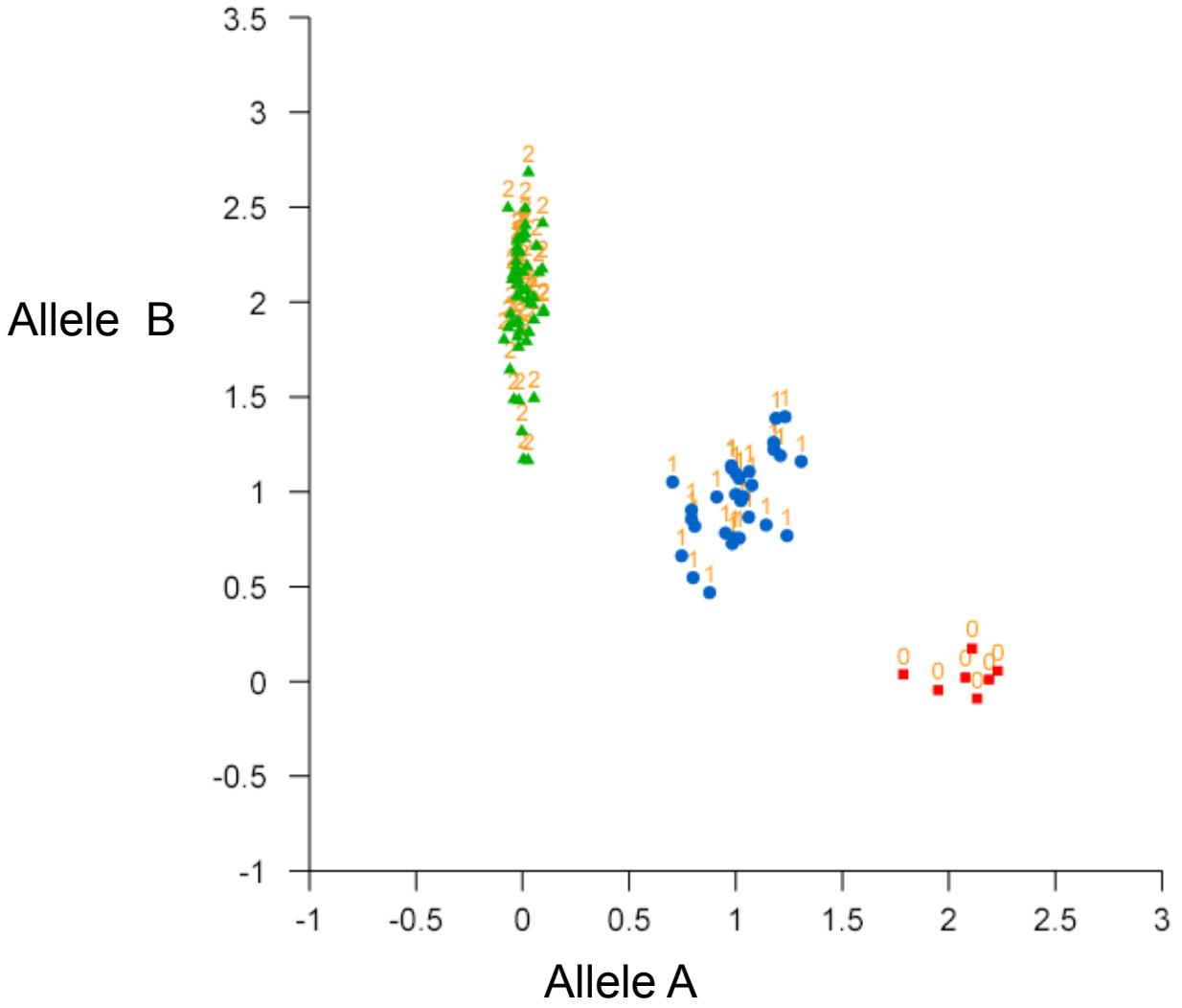
Human Cancer Genome Project

90 humans, 1 SNP (A=0.48)



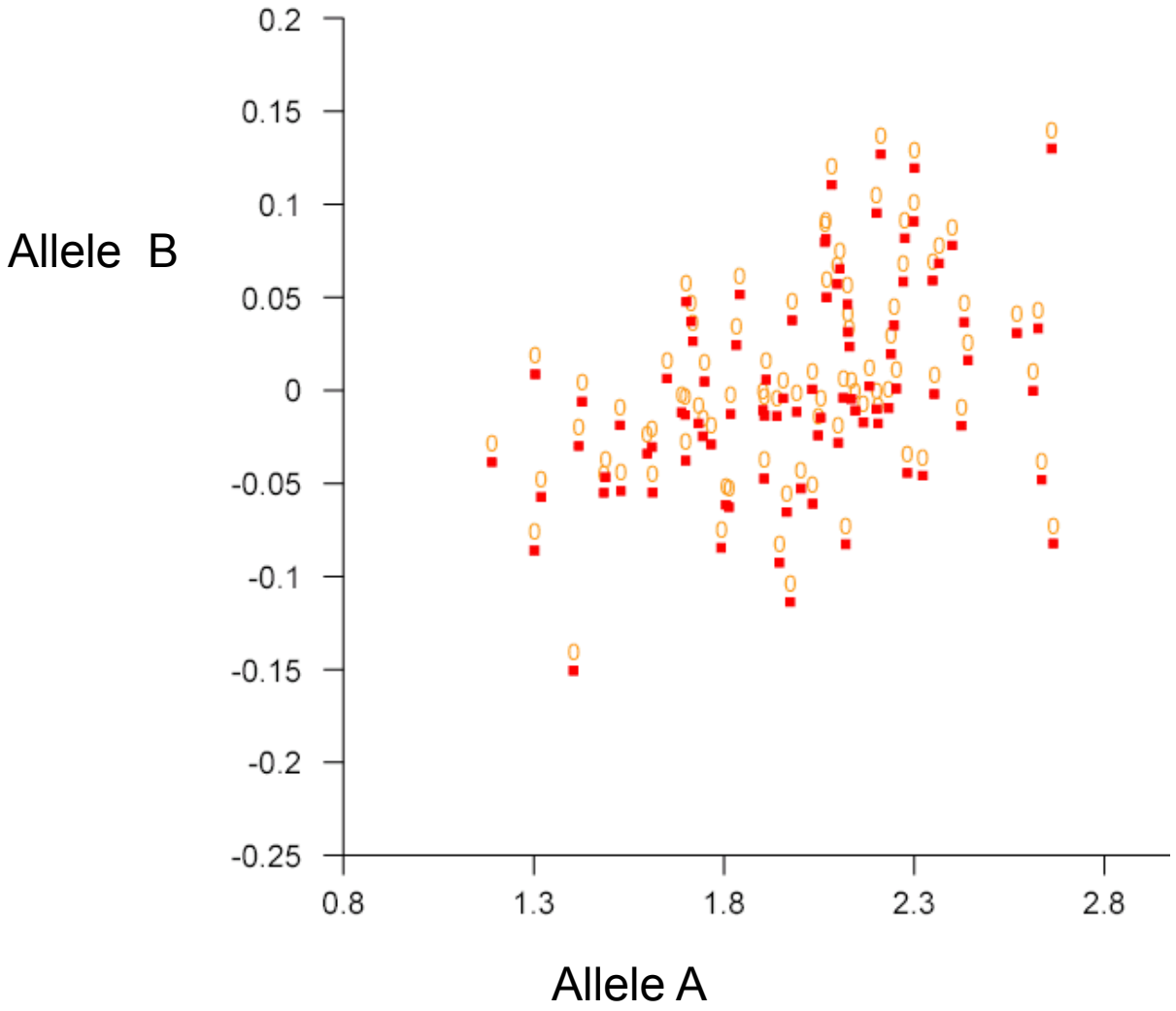
Human Cancer Genome Project

90 humans, 1 SNP (A=0.24)



Human Cancer Genome Project

90 humans, 1 SNP (A=0.96)



Background Correction & Normalization

- Consider a genomic location L and two "similar" nucleotide sequences $s_{L,x}$ and $s_{L,y}$ starting at that location in the two copies of a diploid genomes…
  - E.g., they may differ in one SNP.
  - Let θ_x and θ_y be their respective copy numbers in the whole genome and all copies are selected in the reduced complexity representation. The gene chip contains four probes p_x in $s_{L,x}$; p_y in $s_{L,y}$; $p_{x'}$, $p_{y'}$ not in G.
  - After PCR amplification, we have some $K_x * \theta_x$ amount of DNA that is complementary to the probe p_x, etc. K' (~ K'_x) amount of DNA that is additionally approximately complementary to the probe p_x.

Normalize using a Generalized RMA

$$I' = U - \mu_n - [\alpha \sigma_n^2 - \phi_{N(0,1)}(a'/b')/\Phi_{N(0,1)}(a'/b')] \pounds\{(1 + \beta' B_{\sigma_n}/\Phi_{N(0,1)}(a'/b')\}^{-1} + [b_{\sigma_n}/B_{\sigma_n}])] \pounds\{(1 + \Phi_{N(0,1)}(a'/b')/(\beta' B_{\sigma_n})\}^{-1},$$

- Where $a' = U-\mu_n -\alpha \sigma_n^2$; $b' = \sigma_n$, and
- $b_{\sigma_n} = \sum [I_{i,j} - U + \mu_n] \phi_{N(0,1)}([I_{i,j} - U + \mu_n])$
- $B_{\sigma_n} = \sum \phi_{N(0,1)}([I_{i,j} - U + \mu_n])$

Human Cancer Genome Project

Background Correction & Normalization

- If the probe has an affinity ϕ_x, then the measured intensity is can be expressed as

$$[K_x\ \theta_x + K'] \phi_x + \text{noise}$$
$$= [\theta_x + K'/K_x] \phi'_x + \text{noise}$$

  - **With Exp[$\mu 1 + \varepsilon\ \sigma 1$], a multiplicative logNormal noise, [$\mu 2 + \varepsilon\ \sigma 2$] an additive Gaussian noise, and $\phi'_x = K_x\ \phi_x$ an amplified affinity.**

- A more general model:

$$I_x = [\theta_x + K'/K_x] \phi'_x\ e^{\mu 1 + \varepsilon\ \sigma 1} + \mu 2 + \varepsilon\ \sigma 2$$

Human Cancer Genome Project

Mathematical Model

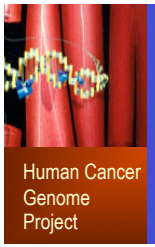
- In particular, we have four values of measured intensities:

$$I_x = [\theta_x \phi'_x + N_x]e^{\mu 1 + \varepsilon \sigma 1} + \mu 2 + \varepsilon \sigma 2$$

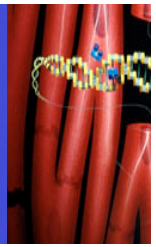
$$I_{x'} = [N_x]\, e^{\mu 1 + \varepsilon \sigma 1} + \mu 2 + \varepsilon \sigma 2$$

$$I_y = [\theta_y \phi'_y + N_y]\, e^{\mu 1 + \varepsilon \sigma 1} + \mu 2 + \varepsilon \sigma 2$$

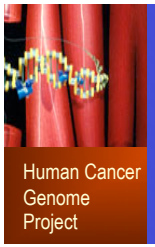
$$I_{y'} = [N_y]\, e^{\mu 1 + \varepsilon \sigma 1} + \mu 2 + \varepsilon \sigma 2$$



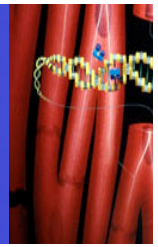
Bioinformatics: Data modeling



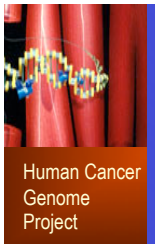
- Good news: For each 25-bp probe, the fluorescent signal increases linearly with the amount of complementary DNA in the sample (up to some limit where it saturates).
- Bad news: The linear scaling and offset differ for each 25-bp probe. Scaling varies by factors of more than 10x.
- Noise : Due to PCR & cross hybridization and measurement noise.



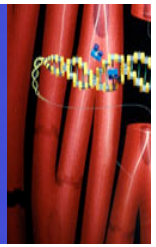
Scaling & Offset differ



- Scaling varies across probes:
  - Each 25-bp sequence has different thermodynamic properties.
- Scaling varies across samples:
  - The scanning laser for different samples may have different levels.
  - The starting DNA concentrations may differ; PCR may amplify differently.
- Offset varies across probes:
  - Different levels of Cross Hybridization with the rest of the Genome.
- Offset varies across samples:
  - Different sample genomes may differ slightly (sample degradation; impurities, etc.)



Linear Model + Noise



i = sample

k = probe in probeset j

PM_{ik} = Observed DNA level

θ_{ik} = True DNA level

$$PM_{ik} = K_i\left(N_k + \theta_{ik}\phi_k\right)e^{\varepsilon\sigma_{ik}} + C_i + \varepsilon'\sigma'_{ik}$$

where

$\varepsilon, \varepsilon'$ are gaussian noise sources

$\sigma_{ik}, \sigma'_{ik}$ are noise scaling factors

Noise minimization

Just estimate θ_{ik} and parameters given PM_{ik} using Maximum Likelihood Estimate (MLE).

This is much simpler if we have only one noise term. We can approximate with a single multiplicative noise term:

$$PM_{ik} \cong K_i\left(N_k + \theta_{ik}\phi_k + F_i\right)e^{\varepsilon\sigma_{ik}} + C_i - K_iF_i$$

Final Data Model

$$A_i(PM_{ik} + B_i) = \left(N_k + \theta_{ik}\phi_k + F_i\right)e^{\varepsilon_{ik}\sigma_{ik}}$$

where

$\sigma_{ik} = s_i t_k$ & θ_{ik} are the same for all probes k in the same probeset j..

The corresponding probability density is :

$$P\left(PM_{ik} \mid \Theta\right) = \frac{e^{-\varepsilon_{ik}{}^2/2}}{\left(PM_{ik} + B_i\right)\sqrt{2\pi\sigma_{ik}{}^2}}$$

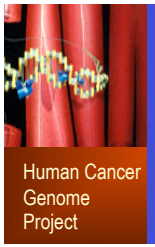
MLE using gradients

Overall log likelihood (no priors):

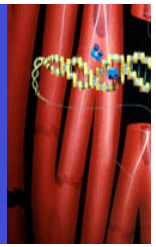
$$L = \sum_{i,k} \log(PM_{ik} + B_i) + \log(s_i t_k) + \log^2\left(\frac{A_i(PM_{ik} + B_i)}{N_k + \theta_{ik}\phi_k + F_i}\right) / \left(2 s_i^2 t_k^2\right)$$

For each parameter $\theta \in \Theta$, gradient update:

$$\theta \rightarrow \theta - \frac{\partial L / \partial \theta}{\partial^2 L / \partial^2 \theta}$$



Data Outliers



- Our data model fails for few data points ("bad probes")
  - Soln (1): Improve the model…
  - Soln (2): Discard the outliers
  - Soln (3): Alternate model for the outliers… Weight the data appropratately.

Human Cancer Genome Project

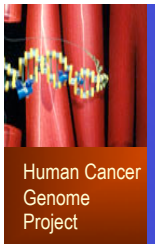
Outlier Model

$$P(PM_{ik}) = w_1 P_1(PM_{ik}) + (1 - w_1) P_2(PM_{ik})$$

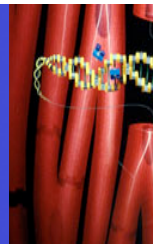
where

$P_2(PM_{ik})$ = Uniform Distribution

w_1 = Prior probability that data is NOT outlier.



Problem with MLE: No unique maxima



The following have no effect on probability :

1. Increase all F_i and decrease all N_k by C.
2. In any probeset j : Increase θ_{ik} by N and decrease N_k by $N\phi_k$
3. Scale all $A_i, N_k, F_i, \theta_{ik}$ by same factor C
4. Scale s_i and unscale t_k by same factor C
5. In any probeset j : Scale ϕ_k and unscale θ_{ik} by same factor C

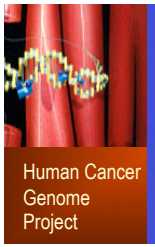
Scaling of MLE estimate

The MLE estimate of θ_{ij} must be rescaled:

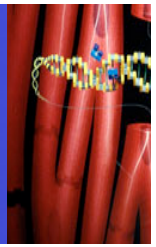
$$\theta_{ij}' = C_j \theta_{ij} + D_j$$

The correct scaling factors C_j, D_j cannot be inferred from the data model.

However we can use priors on the copy number θ_{ij} and the relative frequency of alleles A and B.



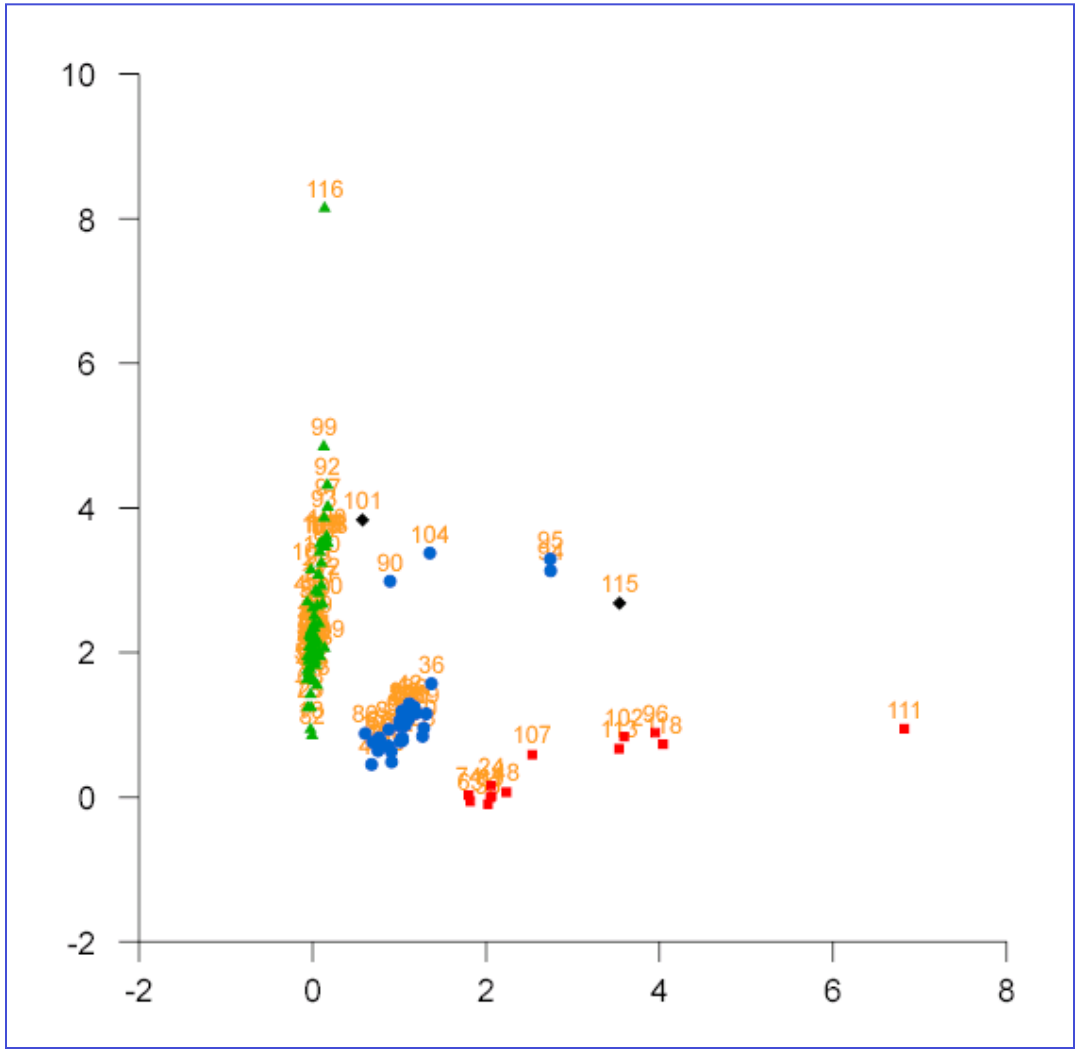
Segmentation to reduce noise



- The true copy number (Allele A+B) is normally 2 and does not vary across the genome, except at a few locations (breakpoints).
- Segmentation can be used to estimate the location of breakpoints and then we can average all estimated copy number values between each pair of breakpoints to reduce noise.

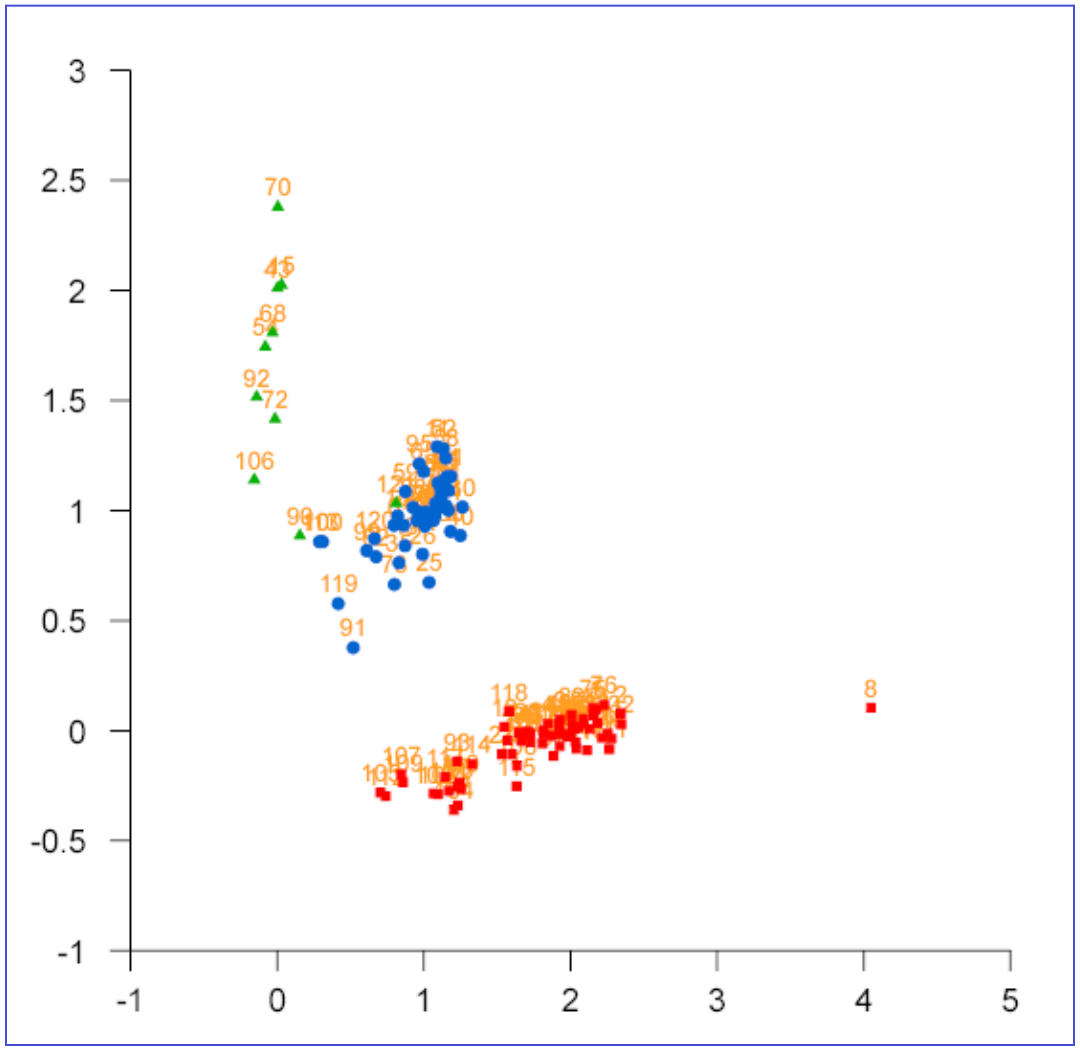
Human Cancer Genome Project

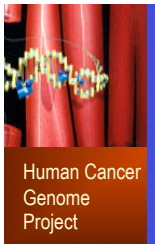
Allelic Frequencies: Cancer & Normal



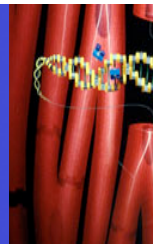
Human Cancer Genome Project

Allelic Frequencies: Cancer & Normal





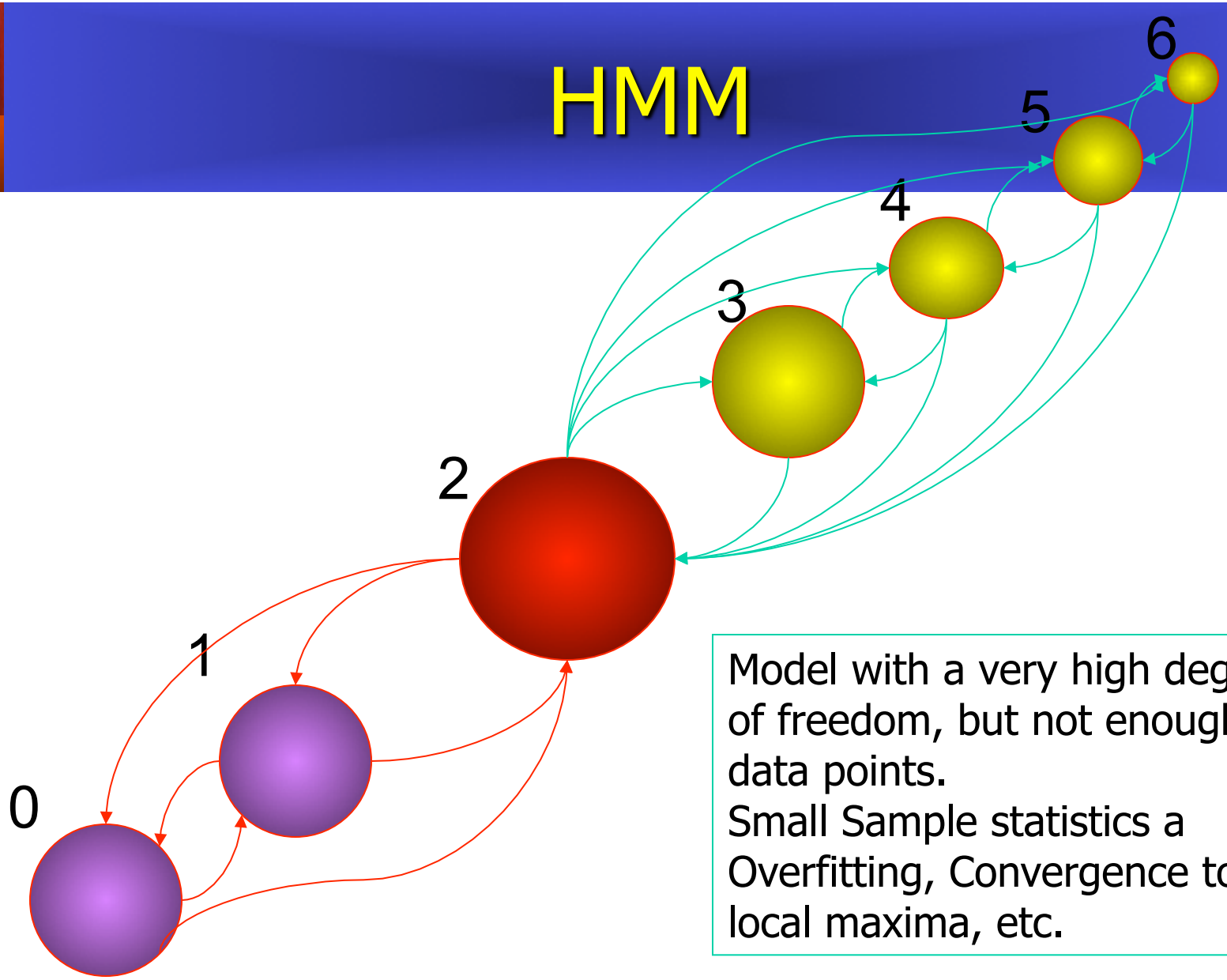
Algorithmic Approaches



- Local Approach
  - Change-point Detection
    - (QSum, KS-Test, Permutation Test)
- Global Approach
  - HMM models
  - Wavelet Decomposition
- Bayesian & Empirical Bayes Approach
  - Generative Models
    - (One- or Multi-level Hierarchical)
  - Maximum A Posteriori

Human Cancer Genome Project

HMM

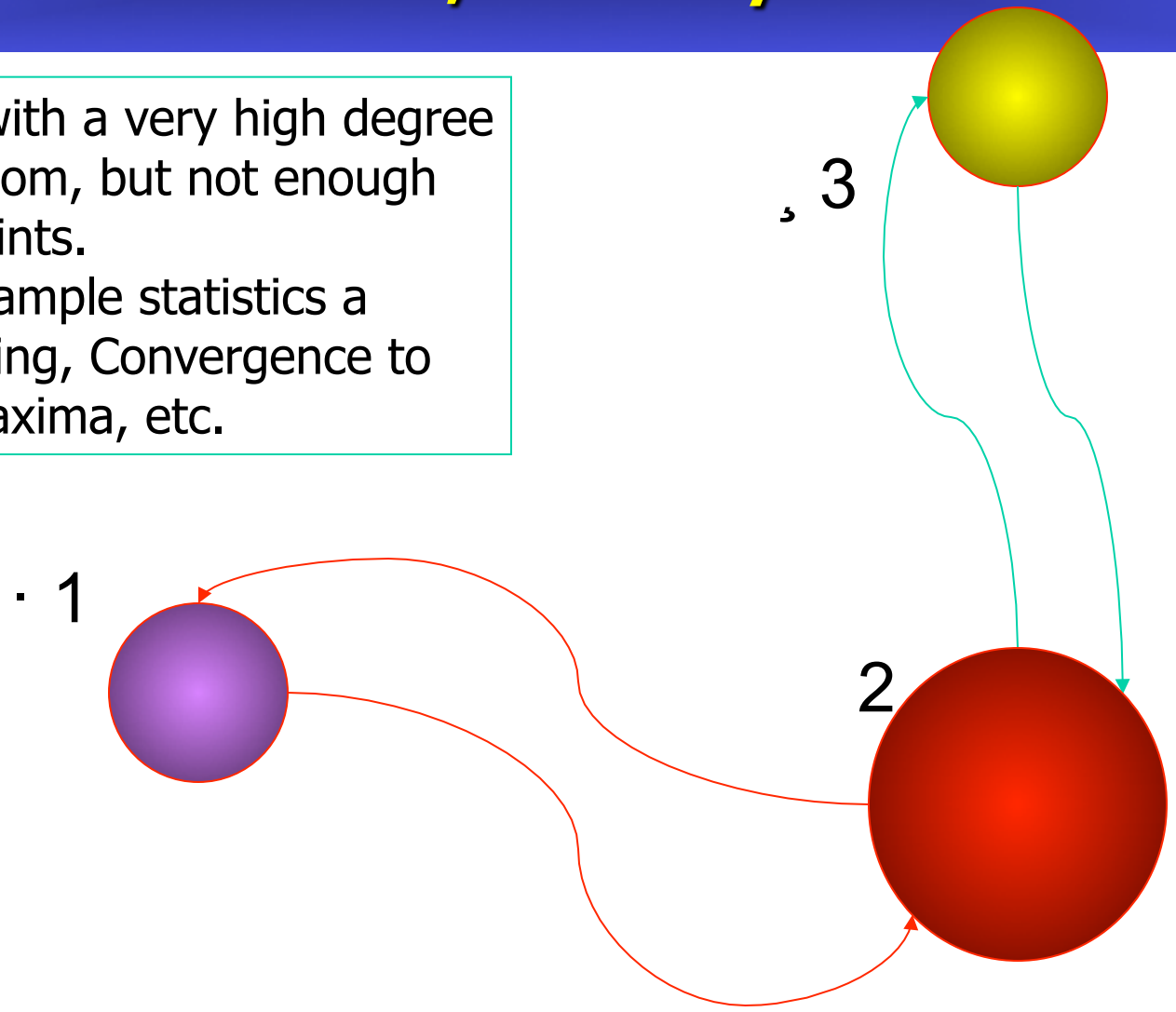


Model with a very high degree of freedom, but not enough data points.
Small Sample statistics a Overfitting, Convergence to local maxima, etc.

Human Cancer Genome Project

HMM, finally…

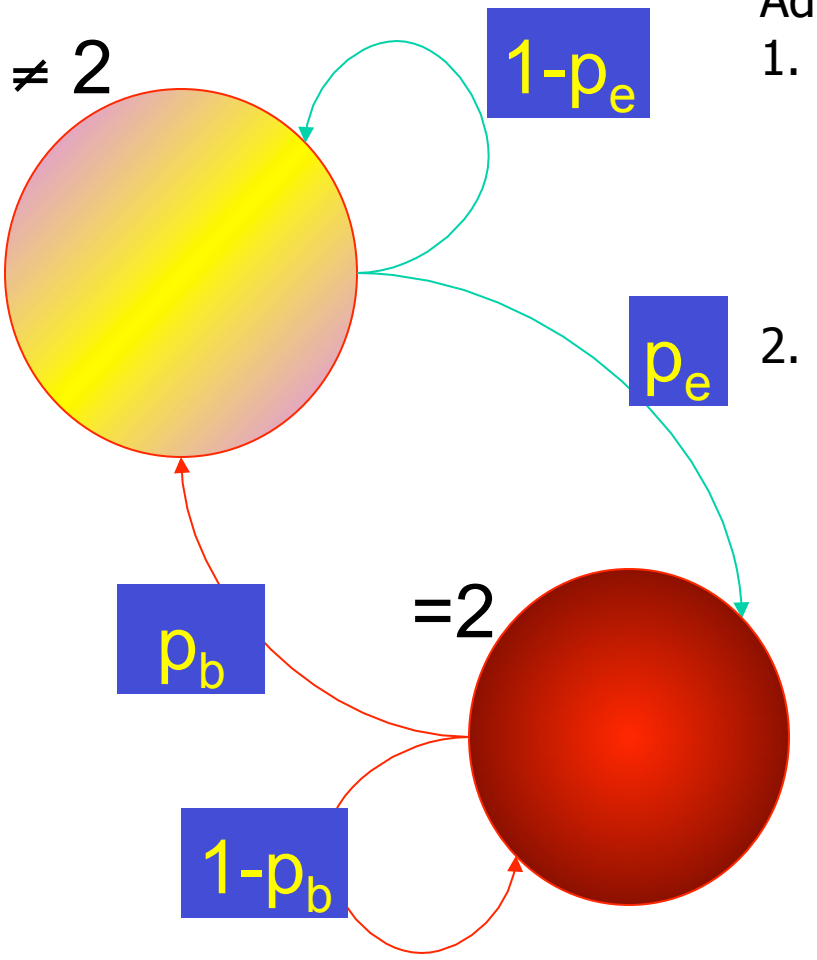
Model with a very high degree of freedom, but not enough data points.
Small Sample statistics a Overfitting, Convergence to local maxima, etc.



HMM, last time

Human Cancer Genome Project

We will simply model the number of break-points by a Poisson process, and lengths of the aberrational segments by an exponential process.
Two parameter model: p_b & p_e



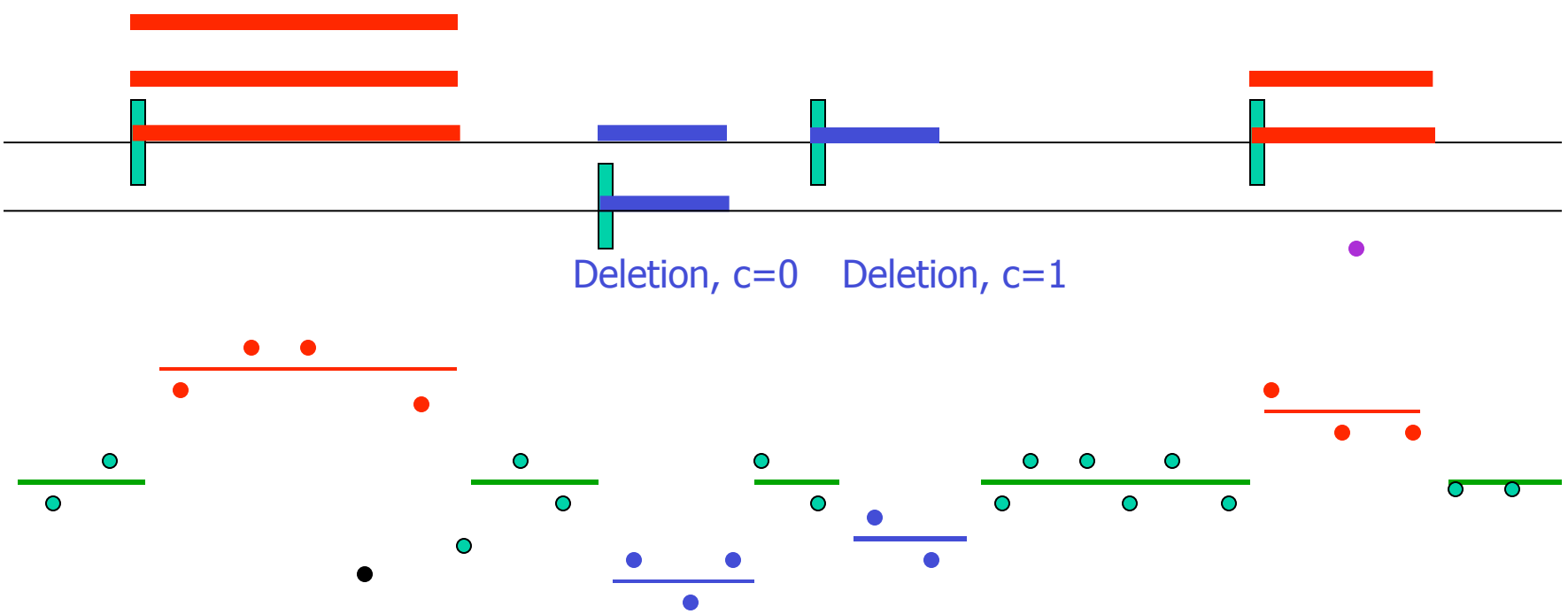
Advantages:

1. Small Number of parameters. Can be optimized by MAP estimator. (EM has difficulties).
2. Easy to model deviation from Markvian properties (e.g., polymorphisms, power-law, Polya's urn like process, local properties of chromosomes, etc.)

Human Cancer Genome Project

Generative Model

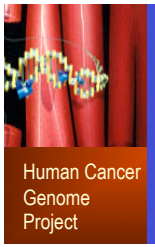
Breakpoints, Poisson, p_b
Segmental Length, Exponential, p_e
Copy number, Empirical Distribution
Noise, Gaussian, μ, σ



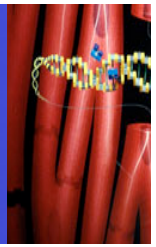
Human Cancer Genome Project

Likelihood Function

- The likelihood function for first n probes:
- $L(< i_1, \mu_1, \ldots, i_k, \mu_k >) = Exp(-p_b\, n)\,(p_b\, n)^k * (2\pi\sigma^2)^{(-n/2)} \prod_{i=1}^{n} Exp[-(v_i - \mu_j)^2/2\sigma^2] * p_e^{(\#global)}(1-p_e)^{(\#local)}$
  - Where $i_k = n$ and i belongs to the j^{th} interval.
  - Maximum A Posteriori algorithm (implemented as a Dynamic Programming Solution) optimizes L to get the best segmentation
- $L(<i^*_1, \mu^*_1, \ldots, i^*_k, \mu^*_k >)$



Dynamic Programming Algorithm



- Generalizes Viterbi and Extends.
- Uses the optimal parameters for the generative model:
- Adds a new interval to the end:
- $h\ i_1, \mu_1, \ldots, i_k, \mu_k\ i \pm h\ i_{k+1}, \mu_{k+1}\ i = h\ i_1, \mu_1, \ldots, i_k, \mu_k, i_{k+1}, \mu_{k+1}\ i$
- Incremental computation of the likelihood function:

$$-\text{Log } L(h\ i_1, \mu_1, \ldots, i_k, \mu_k, i_{k+1}, \mu_{k+1}\ i)$$

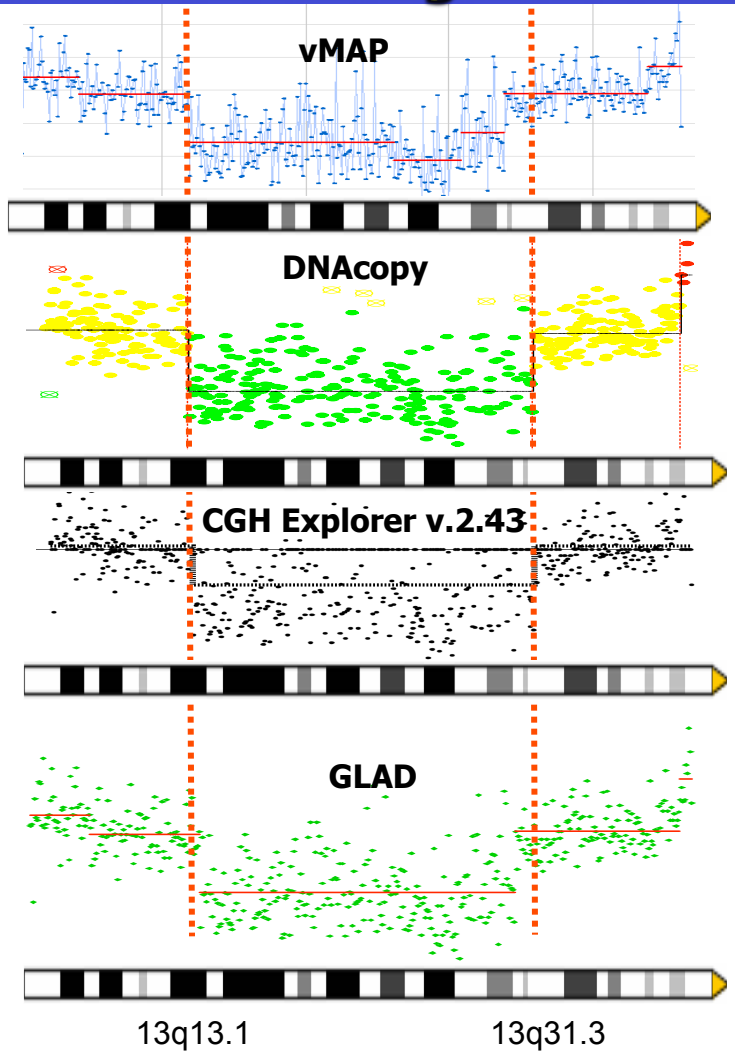
$$= -\text{Log } L(h\ i_1, \mu_1, \ldots, i_k, \mu_k i)$$

$$+ \text{new-res.}/2\sigma^2 - \text{Log}(p_b n) + (i_{k+1} - i_k)\ \text{Log}\ (2\pi\sigma^2)$$

$$- (i_{k+1} - i_k)\ [l_{global}\ \text{Log}\ p_e + l_{local}\ \text{Log}(1 - p_e)]$$

Human Cancer Genome Project

Comparison of chromosome 13 tumor using 4 different segmentation algorithm

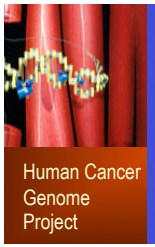


Daruwala et al.
Proc Natl Acad Sci U S A. **2004**

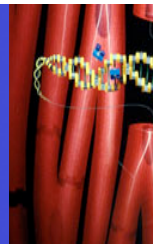
Olshen, AB et al.
*Biostatistic*s **5**: 557-72

Lingjaerde, OC et al.
Bioinformatics **21**: 821-2

Hupe, P et al.
Bioinformatics **20**: 3413-22



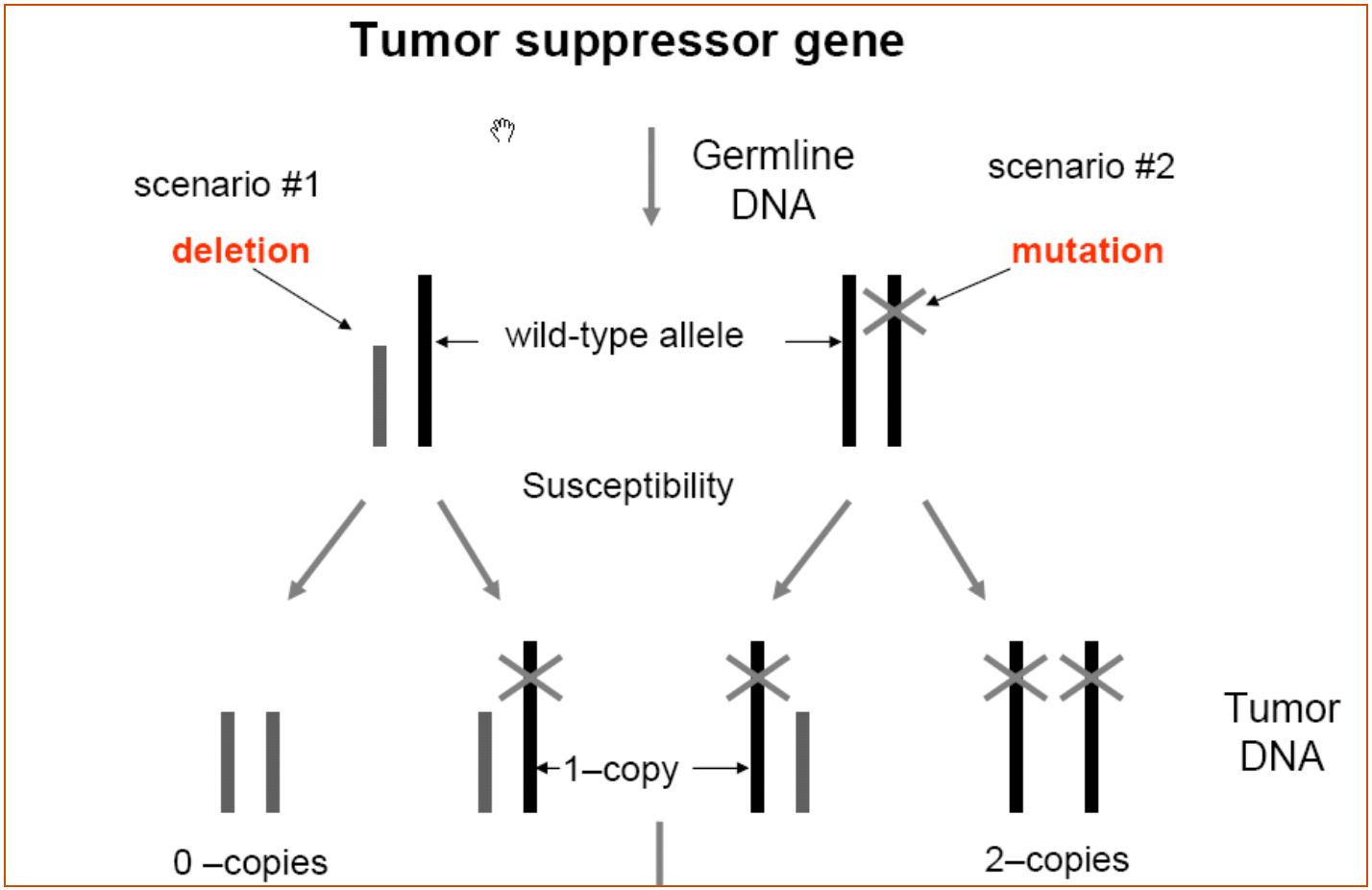
Finding Cancer Genes



- LOH/Deletion Analysis analysis
- Hypothesize a TSG (Tumor Suppressor Gene)
- Score function for each possible genomic region containing the TSG
  - Evolutionary history
  - Interactions
  - Parameters
- This score can be computed using estimation from data and also prior information on how the deletions arise. We use a simple approximation; we assume there is a Poisson process that generates breakpoints along the genome and an Exponential process that models the length of the deletions.

Human Cancer Genome Project

Loss of Heterozygousity



Human Cancer Genome Project

Relative Risk Score

- For an interval I (set of consecutive probes) we define a multipoint score quantifying the strength of associations between disease and copy number changes in I.

$$RR_I = \ln \frac{P(\text{disease} \mid A)}{P(\text{disease} \mid \overline{A})} = \ln \left(\frac{P(A \mid \text{disease})}{P(\overline{A} \mid \text{disease})} \times \frac{P(\overline{A})}{P(A)} \right)$$

$$= \ln \frac{P(A \mid \text{disease})}{P(\overline{A} \mid \text{disease})} - \ln \frac{P(A)}{P(\overline{A})}$$

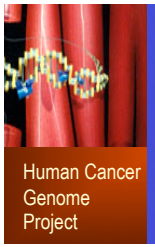
where A is the event "I amplified" (for oncogenes) and "I deleted" (for tumor suppressor genes).

Relative Risk Score (cont)

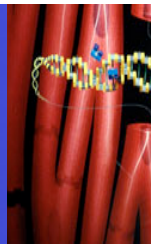
- The first part can be estimated from data:

$$\frac{P(\mathrm{A} \mid \text{disease})}{P(\overline{\mathrm{A}} \mid \text{disease})} = \frac{n_{\mathrm{A}}}{n_{\overline{\mathrm{A}}}}$$

- The second part depends on the marginal probability of amplification (for oncogenes) and deletion (for tumor suppressor genes)



Relative Risk Score: Marginal



- In order to compute the marginal, we rely on the generative model assumed to have produced the data, as follows:
  - Breakpoints occur as a Poisson process at a certain rate μ_a, μ_d
  - At each of these breakpoints, there is an amplification/ deletion with length distributed as an Exponential random variable with parameter λ_a, λ_d

Human Cancer Genome Project

Relative Risk Score: Marginal

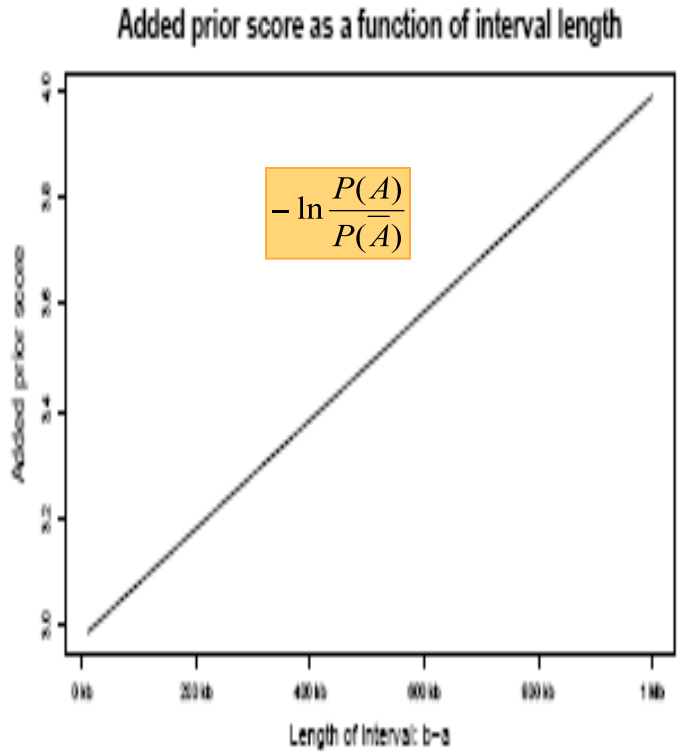
- Assuming the generative process above, we can compute the second part. It depends on the parameters of the Poisson and Exponential random variables.These parameters are estimated from data.

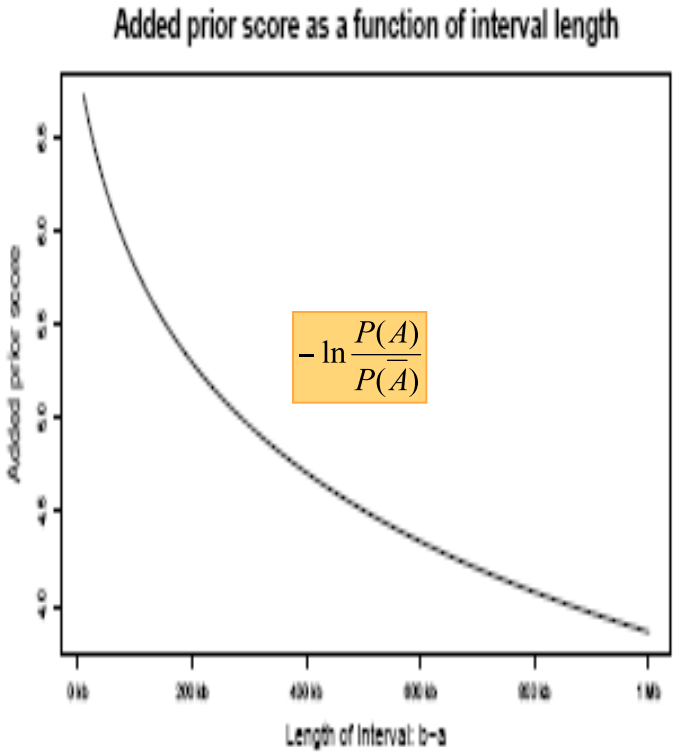
$$1.\ P([a,b]\ \text{amplified}) = 1 - e^{-\mu a e^{-\lambda(b-a)}\frac{1-e^{-\lambda a}}{2\lambda a}}\ e^{-\mu(G-b)e^{-\lambda(b-a)}\frac{1-e^{-\mu(G-b)}}{2\mu(G-b)}}$$

$$2.\ P([a,b]\ \text{deleted}) = 1 - e^{-\mu(b-a)}e^{-\mu a\frac{1-e^{-\lambda a}}{2\lambda a}}\ e^{-\mu(G-b)\frac{1-e^{-\mu(G-b)}}{2\mu(G-b)}}$$

Human Cancer Genome Project

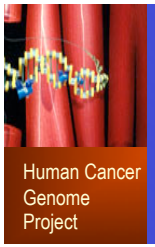
Prior Score



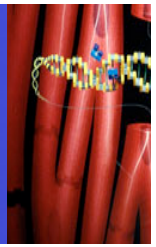


Oncogene

TSG



Finding the Cancer Genes



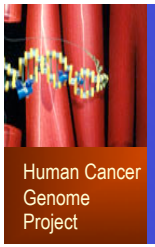
- So far we have shown how to compute the score for a certain genomic intervals

  Intervals with high scores are interesting

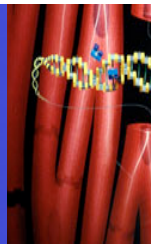
  - Given a larger genomic region, for example a chromosome arm, we compute the scores for all possible intervals up to a certain length
  - The maximum scoring interval in a region is the most likely location for a cancer gene
- We propose two methods to estimate the location of possible cancer genes in this region:

  The Max method &

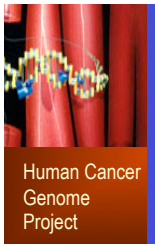
  The LR (left-right) method



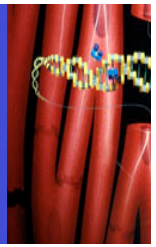
High Scoring Intervals



- High scoring intervals are obvious candidates for cancer genes.
  - We assign significance based on the estimated number of breakpoints in a genomic region with high score.
  - We obtain an approximate p-value using results from scan statistics.



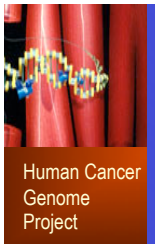
Significance Testing



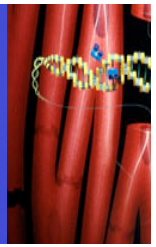
- We now know how to estimate the most likely location of a cancer gene in a genomic region of interest. Let us say the interval is I_{max}

Is this finding statistically significant?

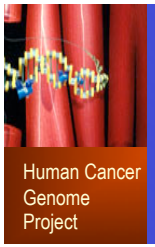
- We rely on an empirical way to compute an approximate p-value



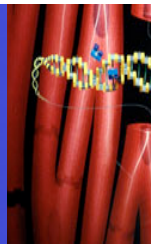
Significance Testing (for TSG)



- The p-value is estimated from the observed distribution of breakpoints along the chromosome
  - Intuitively, in the null hypothesis, which assumes that no tumor suppressor gene resides on the chromosome, the breakpoints are expected to be uniformly distributed
  - However if indeed I_{tsg} is a tumor suppressor gene, then its neighborhood should contain an unusually large number of breakpoints, signifying a region with many deletions



Scan Statistics



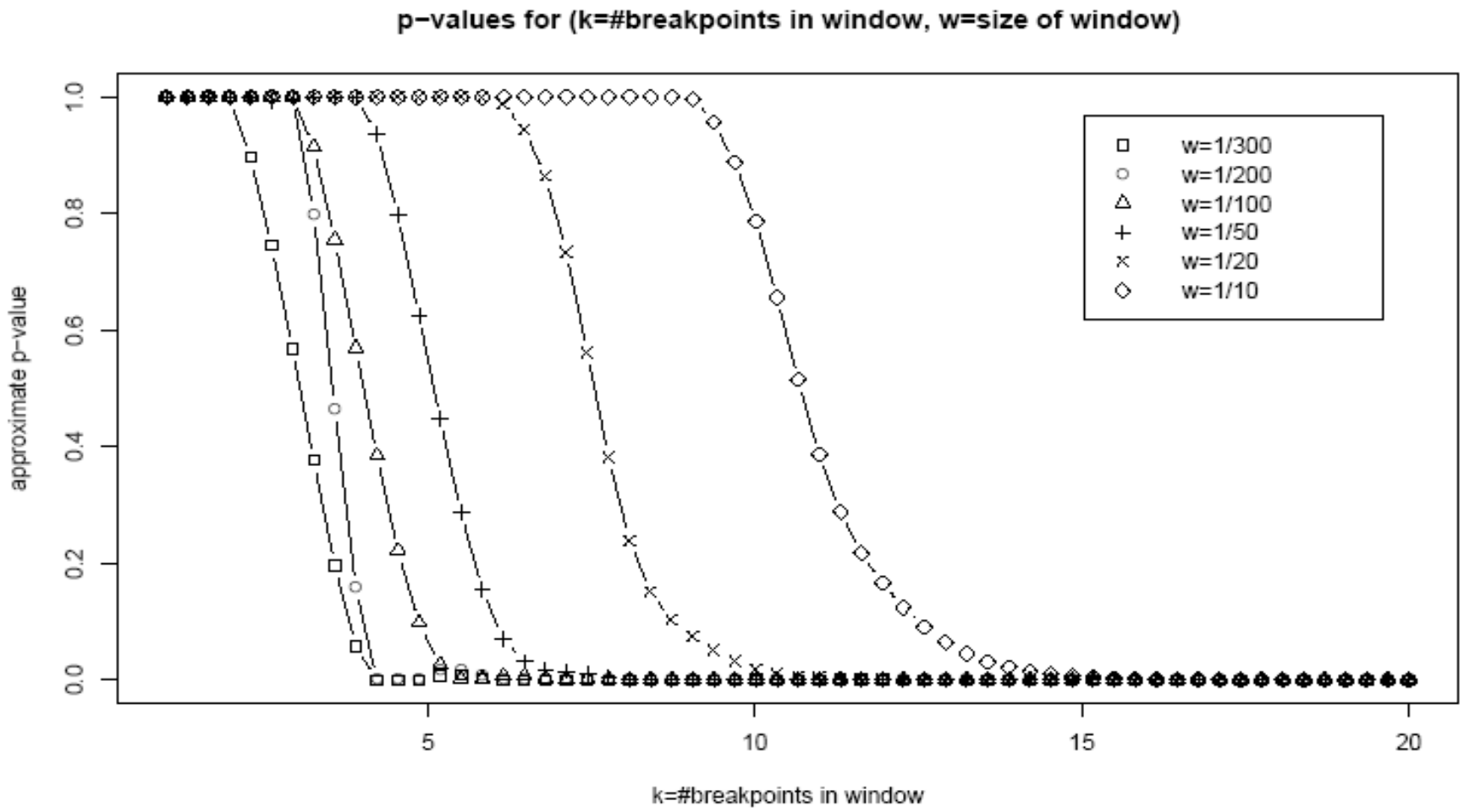
- If N is the total number of breakpoints on the chromosome and k is the number of breakpoints in I_{tsg}, then we can compute the probability of observing k out of N breakpoints in a window of length $|I_{tsg}|(=w)$ if these breakpoints are uniformly distributed ¸ p-value

$$P(S_w \geq k) \approx (kw^{-1} - N - 1)b(k; N, w) + 2G_b(k; N, w) \text{ where}$$

$$b(k; N, w) = C(n, k)w^k (1-w)^{N-k} \text{ and } G_b(k; N, w) = \sum_{i=k}^{N} b(i; N, w)$$

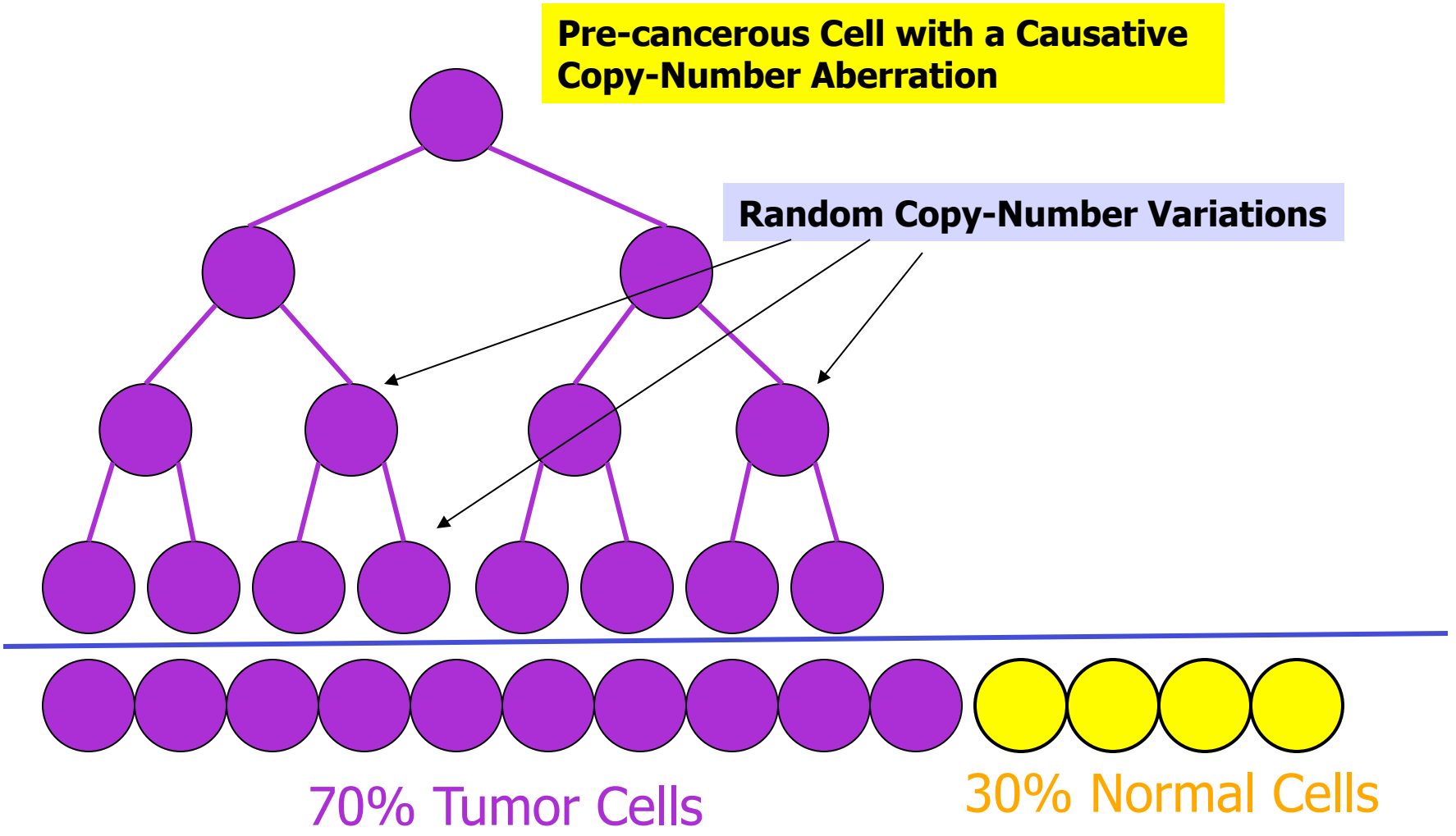
Human Cancer Genome Project

Scan Statistics



Human Cancer Genome Project

Simulation Model

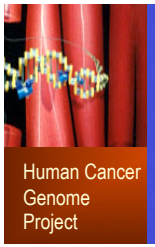


Human Cancer Genome Project

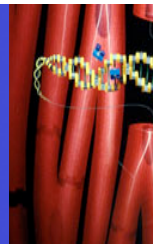
Results – Simulated Data

- We simulated data on diseased people assuming different scenarios. We vary the relative proportions of types of patients in a sample; some patients are diseased because of homozygous deletions of the tumor suppressor gene (a), other because of hemizygous deletions (b) and the rest are diseased because of other causes (c).
- We measure the performance using the Jaccard measure of overlap between the estimated TSG and the true position:

$$J(E,T) = \frac{|E \cap T|}{|E \cup T|}$$



Results



Model	$p_{\text{homozygous}}$	$p_{\text{hemizygous}}$	p_{sporadic}
1	100%	0%	0%
2	50%	50%	0%
3	0%	100%	0%
4	50%	0%	50%
5	25%	25%	50%
6	0%	50%	50%

Table 1: Six simulated models.

Model	Jaccard M. LR	Jaccard M. Max	Sensitivity LR	Sensitivity Max
1	0.82 ± 0.11	0.72 ± 0.23	0.80 ± 0.08	0.79 ± 0.10
2	0.84 ± 0.12	0.67 ± 0.24	0.69 ± 0.10	0.67 ± 0.13
3	0.84 ± 0.10	0.62 ± 0.30	0.56 ± 0.11	0.54 ± 0.13
4	0.74 ± 0.15	0.23 ± 0.19	0.80 ± 0.14	0.69 ± 0.12
5	0.73 ± 0.16	0.33 ± 0.25	0.69 ± 0.12	0.59 ± 0.16
6	0.74 ± 0.17	0.26 ± 0.25	0.54 ± 0.12	0.46 ± 0.12

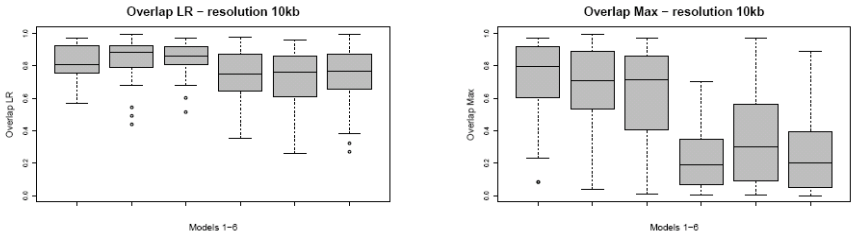
Table 2: Overlap between true location and estimated location of the TSG and the resulting sensitivity. Average inter-marker distance is 10 Kb.

Model	Jaccard M. LR	Jaccard M. Max	Sensitivity LR	Sensitivity Max
1	0.70 ± 0.15	0.44 ± 0.27	0.59 ± 0.16	0.56 ± 0.16
2	0.70 ± 0.19	0.38 ± 0.30	0.46 ± 0.14	0.43 ± 0.15
3	0.68 ± 0.20	0.43 ± 0.30	0.38 ± 0.14	0.34 ± 0.16
4	0.60 ± 0.21	0.25 ± 0.21	0.60 ± 0.18	0.55 ± 0.15
5	0.65 ± 0.20	0.24 ± 0.22	0.46 ± 0.15	0.40 ± 0.14
6	0.58 ± 0.28	0.27 ± 0.28	0.37 ± 0.15	0.33 ± 0.14

Table 3: Overlap between true location and estimated location of the TSG and the resulting sensitivity. Average inter-marker distance is 20 Kb.

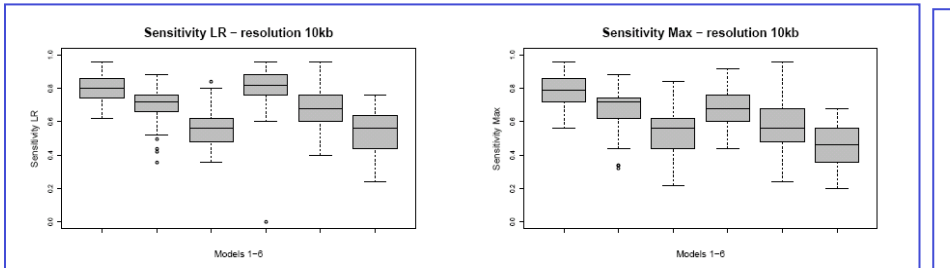
Human Cancer Genome Project

Results



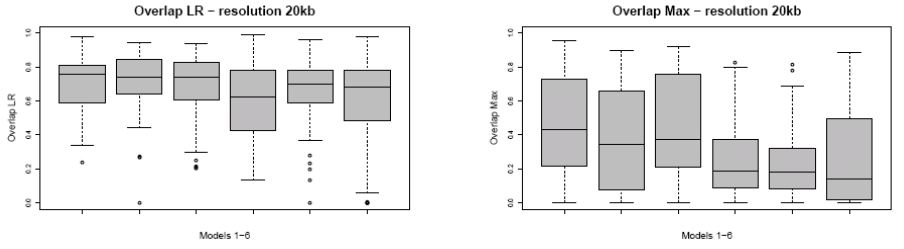
(a) (b)

Figure 2: Boxplots of the Jaccard measure of overlap for each of the six models. Average inter-marker distance is 10 Kb. (a) LR, (b) Max.



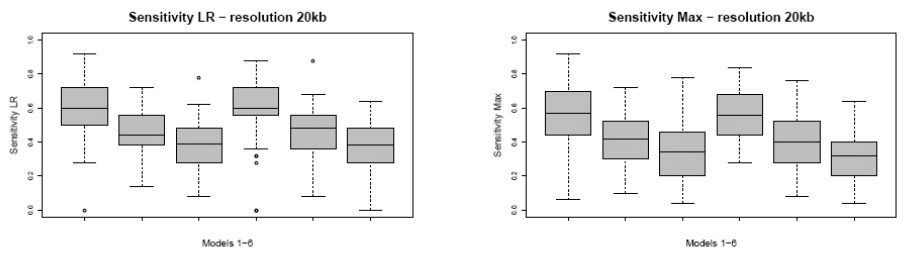
(a) (b)

Figure 3: Boxplots of the sensitivity measure for each of the six models. Average inter-marker distance is 10 Kb. (a) LR, (b) Max.



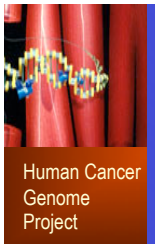
(a) (b)

Figure 4: Boxplots of the Jaccard measure of overlap for each of the six models. Average inter-marker distance is 20 Kb. (a) LR, (b) Max.

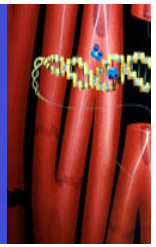


(a) (b)

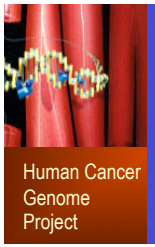
Figure 5: Boxplots of the sensitivity measure for each of the six models. Average inter-marker distance is 20 Kb. (a) LR, (b) Max.



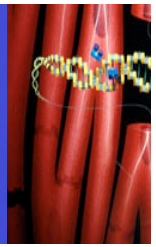
Lung Cancer Dataset



- Dataset of Zhao et al. 2005
  - 70 lung tumors
  - DNA copy number changes across 115,000 SNPs
- First, we infer the copy-number values at these probes and decide which of them are deleted or amplified…



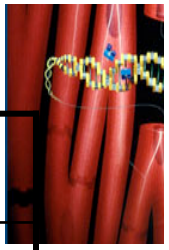
Results



- Most of the regions detected have been previously reported as implicated in lung cancer (e.g. 5q21, 14q11).
- Most significantly, some of the intervals found overlap some good candidate genes, that may play a role in lung cancer (e.g. MAGI3, HDAC11, PLCB1).
- Also, the regions 3q25 and 9p23 have been found for the first time to be homozygously deleted by Zhao et al. (2005).

TSG

Chromosome	Comments
1p13.2	*MAGI3*
3p25.1	*HDAC11*
3q25.1	Homozygous Del
4q34.1	Del Lung Cancer
5q14.1	
5q21.3	Del Lung Cancer
16q24	*CDH13*
17q21	*BRCA1, HDAC5*
19p13.3	*LKB1*
20p12	*PLCB1*
21q21.2	Del Lung Cancer



Onco-gene

Chromosome	Comments
3q28	Over-expression in LC
5p15.3	*LOC389267 (*similar to *MUC4)*
6p22.3	
8q24	*PVT1/MYC*
11p15	*OR51A2*
12p11	Amplification in Zhao et al. (2005)
20q11.23	Amplification in Zhao et al. (2004)