Computational Systems Biology

… Biology X – Lecture 11…

Bud Mishra

Professor of Computer Science, Mathematics, & Cell Biology

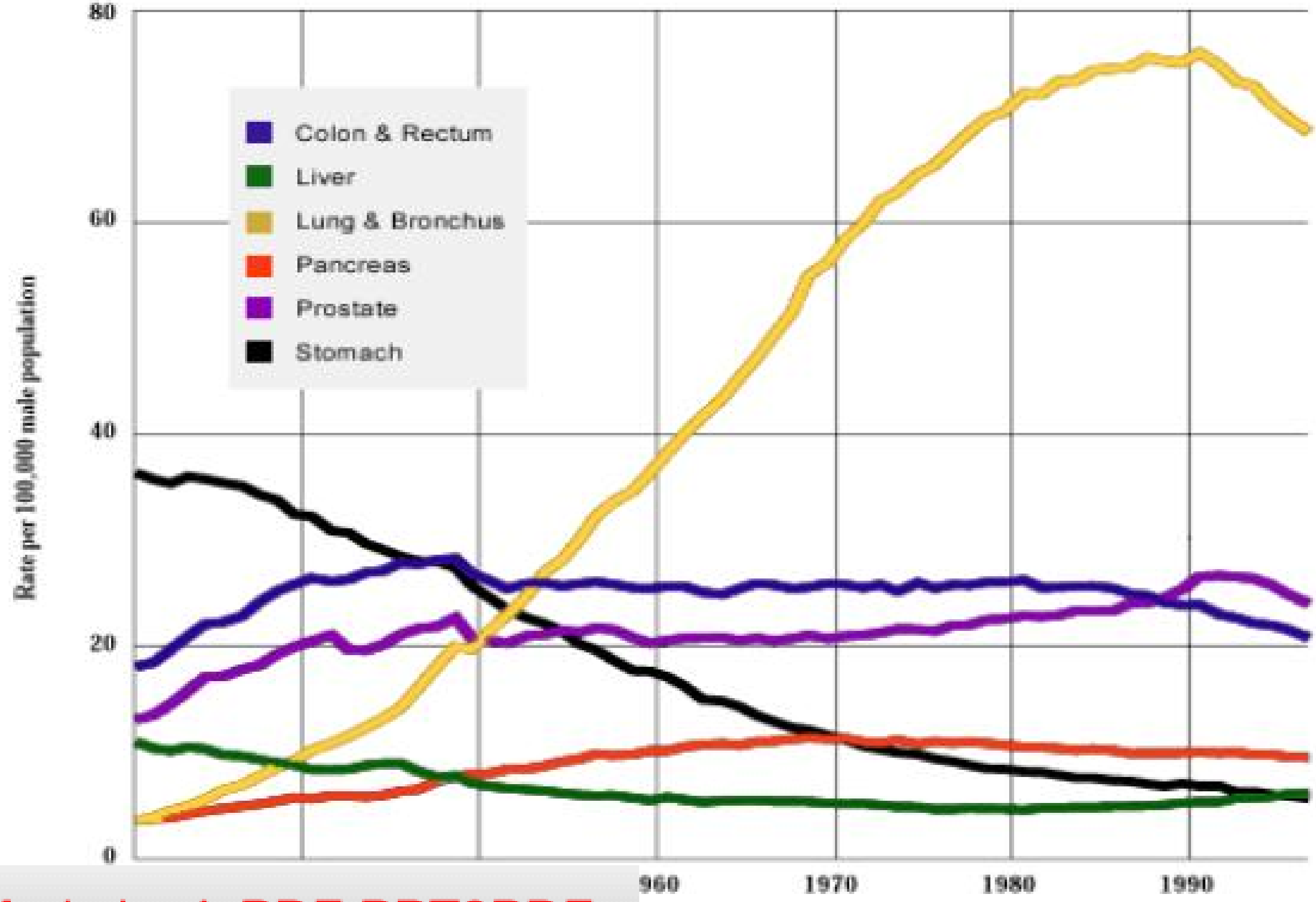
What is Cancer?

- **Cancer develops when cells in a part of the body begin to grow out of control.**
  - Although there are many kinds of cancer, they all start because of out-of-control growth of abnormal cells.
- **Normal body cells grow, divide, and die in an orderly fashion. Cancer cells do not follow this order.**
  - During the early years of a person's life, normal cells divide more rapidly until the person becomes an adult.
  - After that, cells in most parts of the body divide only to replace worn-out or dying cells and to repair injuries

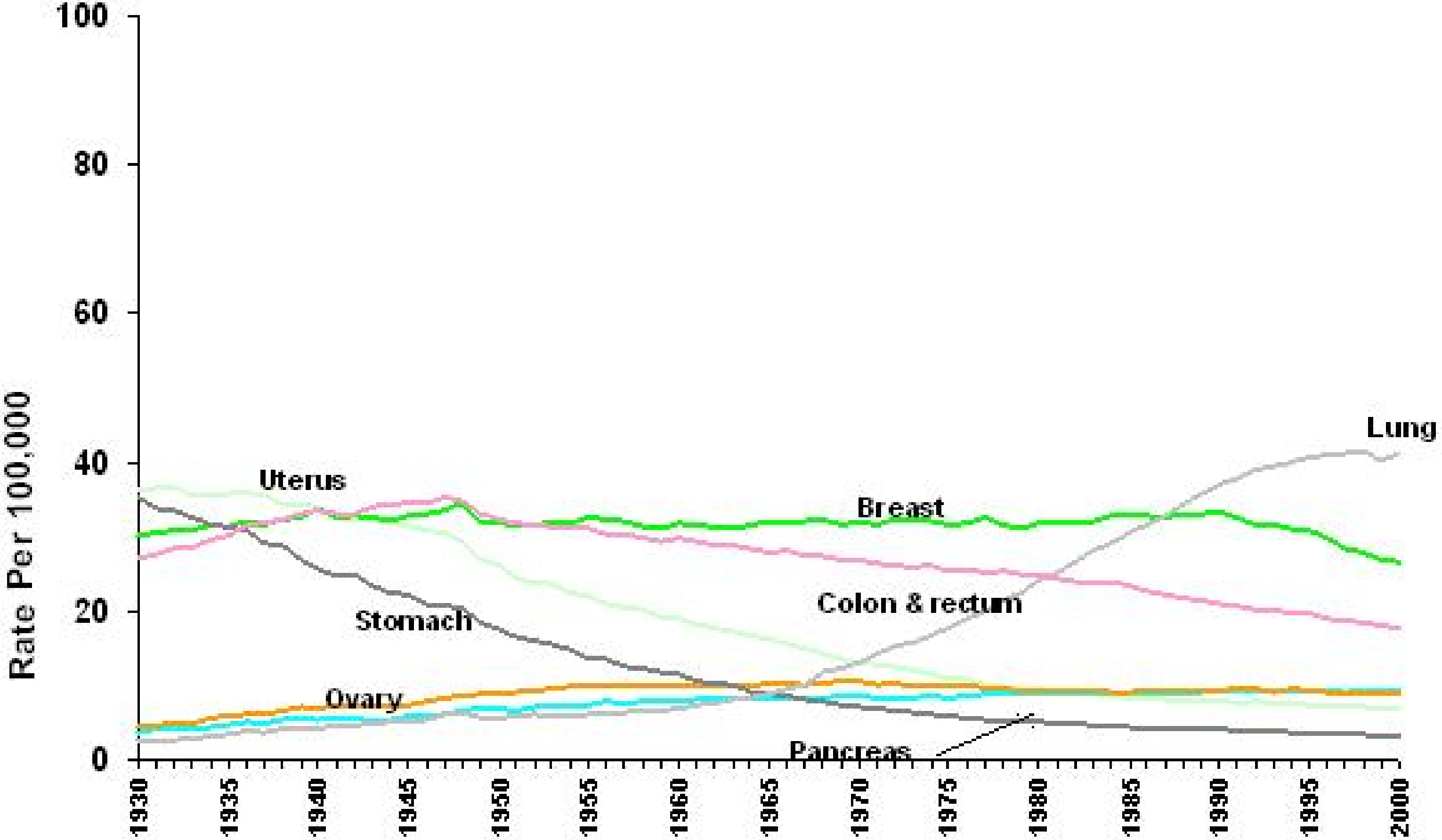
What is Cancer?

- Unlike normal cells, cancer cells continue to grow and divide, and instead of dying, they outlive normal cells and continue to form new abnormal cells.
- Cancer cells develop because of damage to DNA.
  - Most of the time when DNA becomes damaged the body is able to repair it. In cancer cells, the damaged DNA is not repaired.
  - People can inherit damaged DNA, which accounts for inherited cancers. Many times though, a person's DNA becomes damaged by exposure to something in the environment, like smoking.

Male Cancer Death Rates



Cancer Death Rates, for Women



*Age-adjusted to the 2000 US standard population.

es 1960-2000, US Mortality Volumes 1930-1959,

ers for Disease Control and Prevention, 2003.

What is the molecular basis of cancer?

- Cancer is a genetic disease.
  - Mutations in genes result in altered proteins
    - During cell division
    - External agents
    - Random event
  - Most cancers result from mutations in somatic cells
  - Some cancers are caused by mutations in germline cells

Theories of cancer genesis

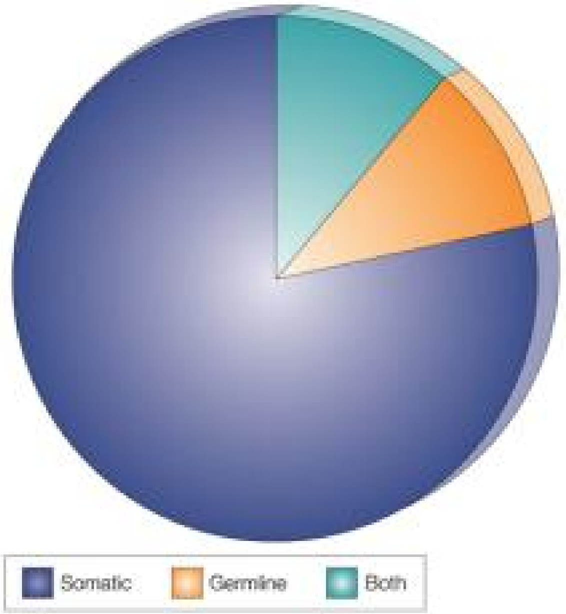
- Standard Dogma
  - Proto-oncogenes (Ras – melanoma)
  - Tumor suppressor genes (p53 – various cancers)
- Modified Dogma
  - Mutation in a DNA repair gene leads to the accumulation of unrepaired mutations (Loeb, 1974) (xeroderma pigmentosum)
- Early-Instability Theory
  - Master genes required for adequate cell reproduction are disabled, resulting in aneuploidy (Philadelphia chromosome)

Types of Cancers

- Based on cells in which they originate
  - Carcinomas (skin, digestive tract)
  - Leukemia (blood forming tissue)
  - Melanoma (epidermis)
  - Sarcoma (connective tissue)
  - Teratoma (germ cells)

Some oncogene statistics

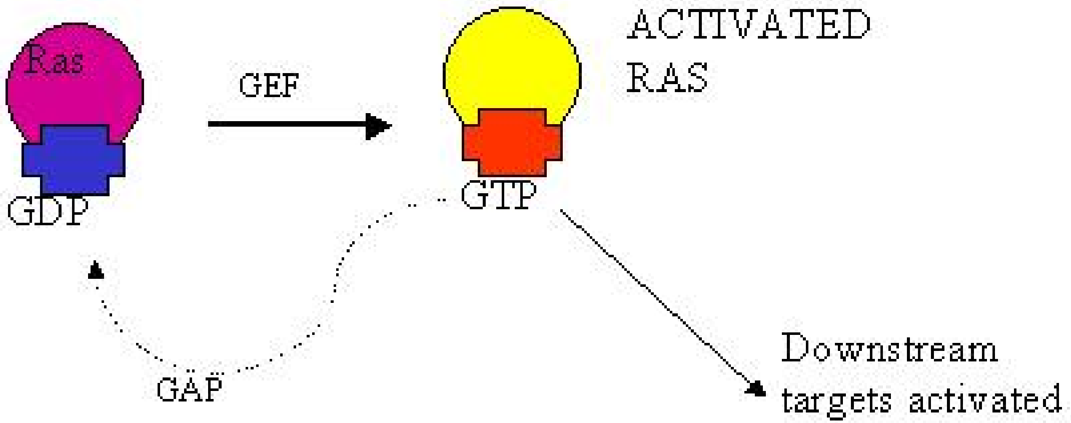
- Approx. 291 genes (1% of the genome)



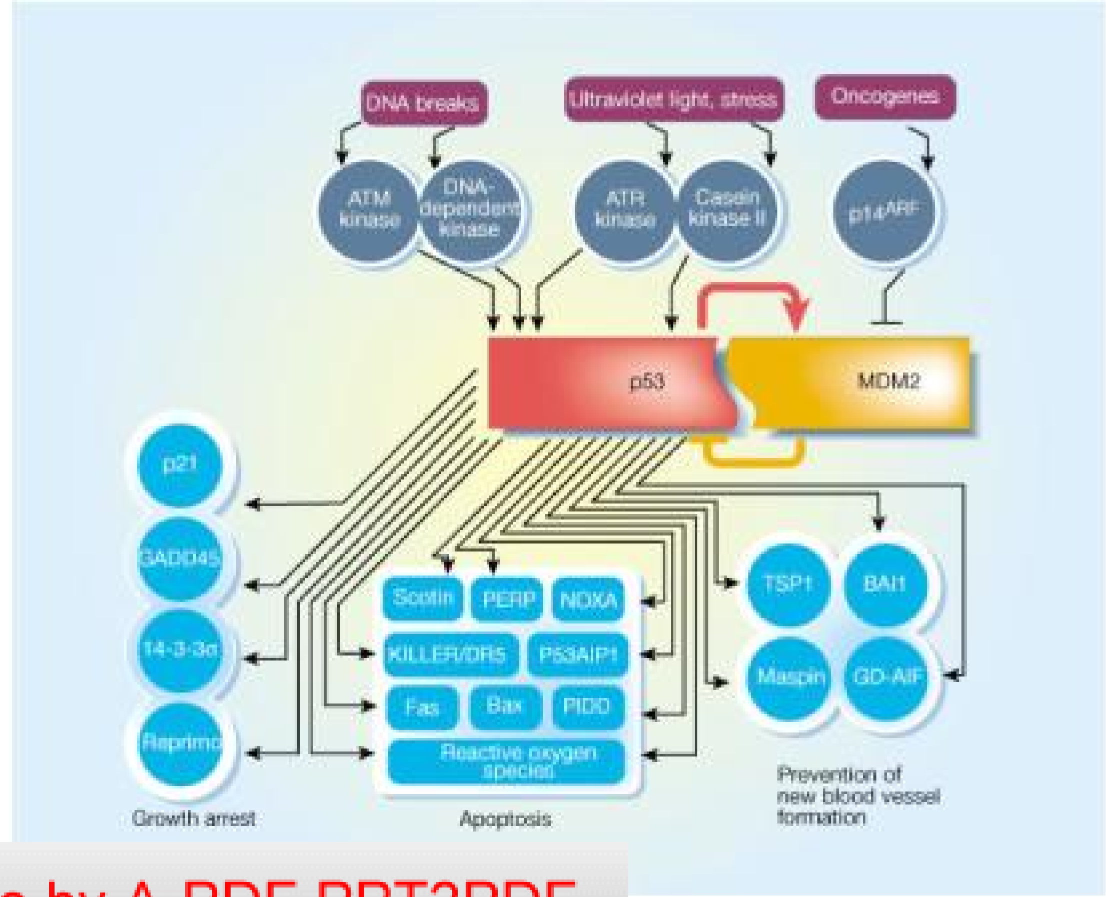
ture Reviews | Cancer

ras oncogene

- Point mutation in codon 12: GGC->GTC



p53 network


DNA breaks
Ultraviolet light, stress
Oncogenes
ATM kinase
DNA-dependent kinase
ATR kinase
Casein kinase II
p14ARF
p53
MDM2
p21
GADD45
14-3-3σ
Reprimo
Scotin
PERP
NOXA
KILLER/DR5
P53AIP1
Fas
Bax
PIDD
Reactive oxygen species
TSP1
BAI1
Maspin
GD-AIF
Growth arrest
Apoptosis
Prevention of new blood vessel formation

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

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.
- 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

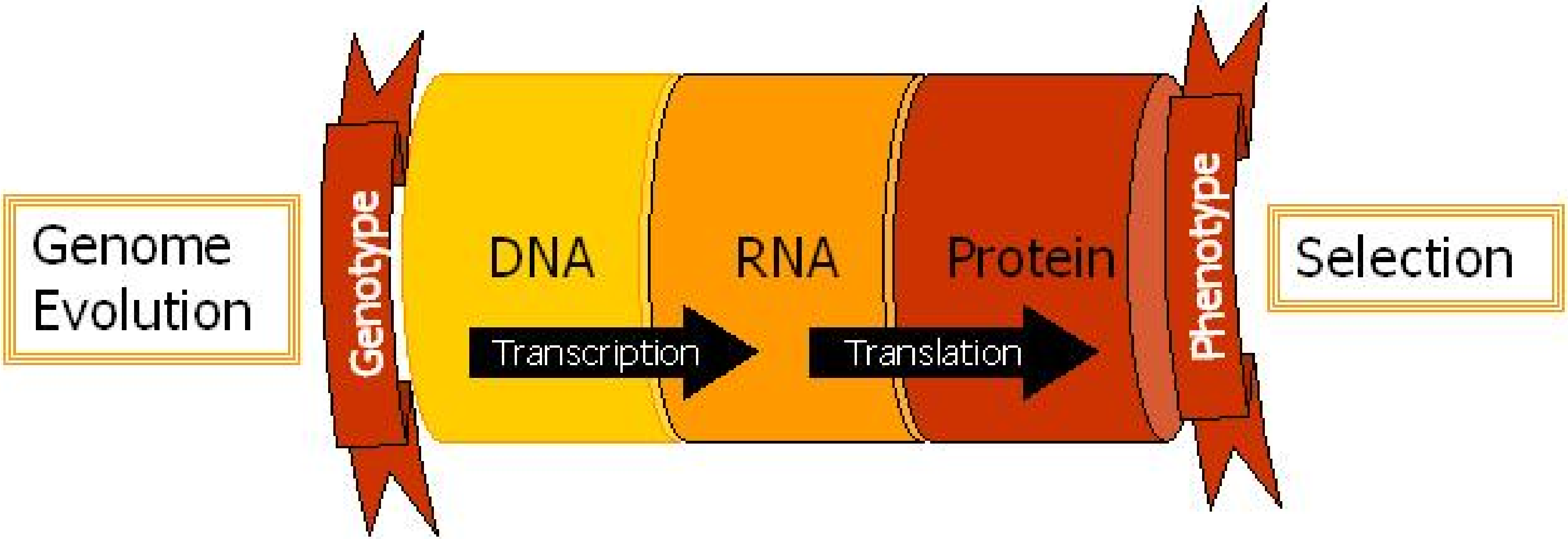
Scientific Foundation for a Human Cancer Genome Project

- Genomic loss and amplification.
  - Consistent association with genomic loss or amplification in many specific regions, indicating that these regions harbor key cancer associated genes
- Gene resequencing.
  - Specific gene classes (such as kinases and phosphatases) in particular cancer types.
- 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.
- Epigenetic changes.
  - Loss of function of tumor suppressor genes by epigenetic modification of the genome — such as DNA methylation and histone modification.

Genomic Basis

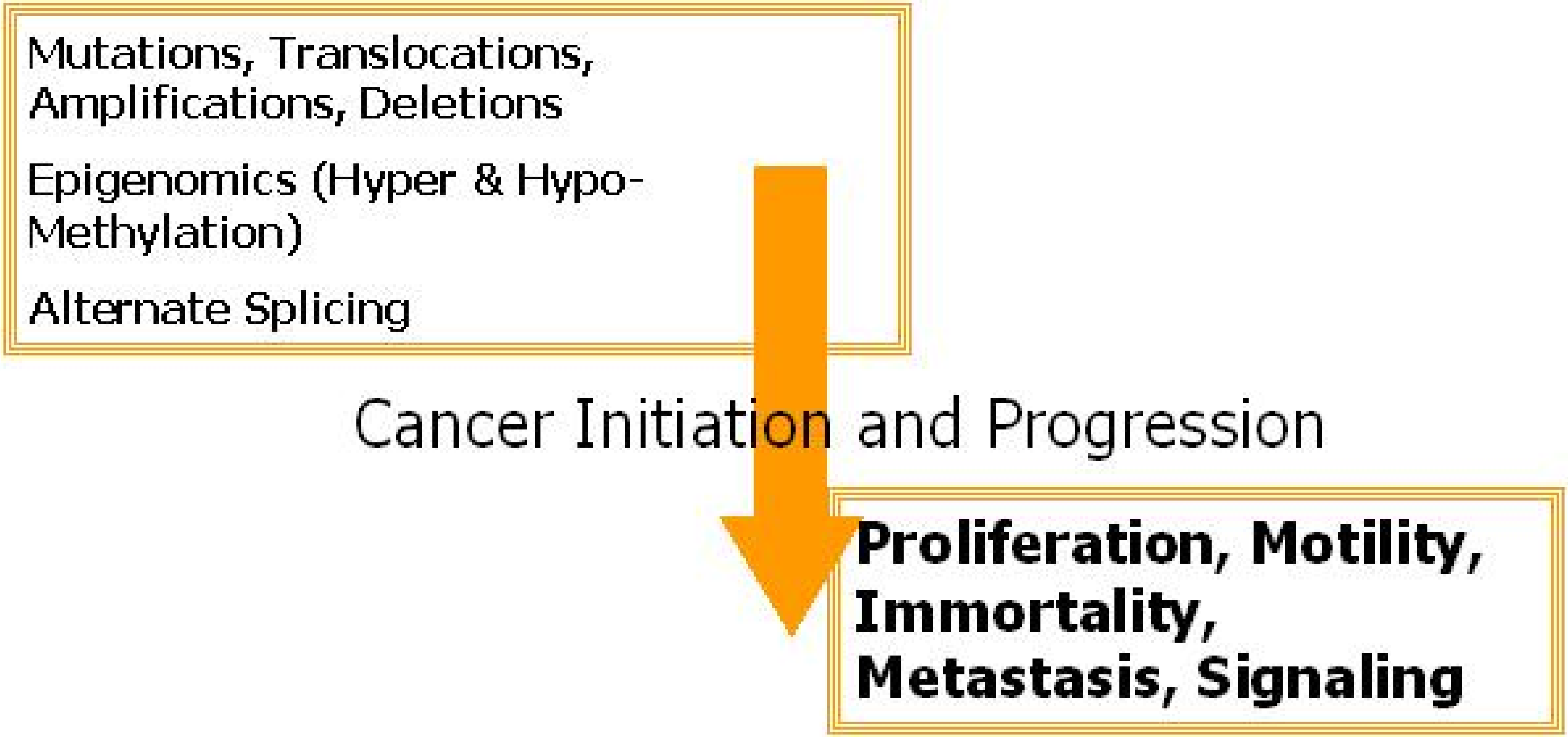
. . .

The New Synthesis



Part-lists, Annotation, Ontologies

Cancer Initiation and Progression



Mutations, Translocations, Amplifications, Deletions

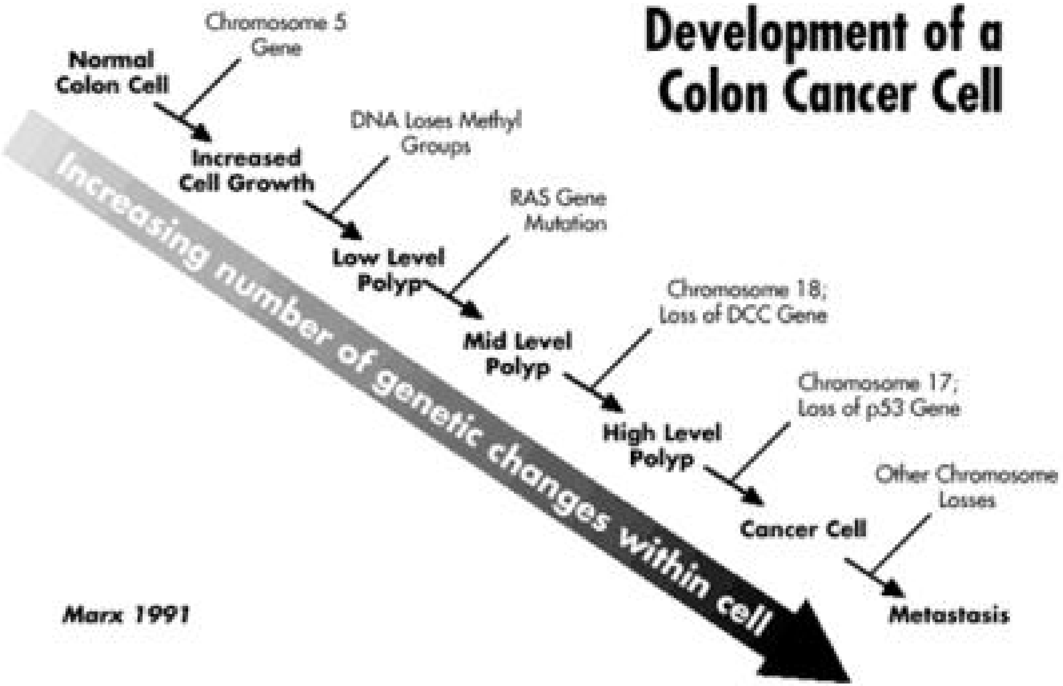
Epigenomics (Hyper & Hypo-Methylation)

Alternate Splicing

Cancer Initiation and Progression

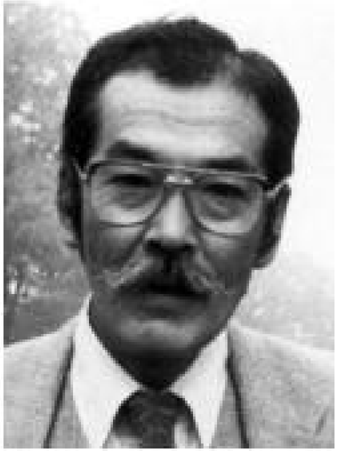
Proliferation, Motility, Immortality, Metastasis, Signaling

Amplifications & Deletions



Human Genome Structure

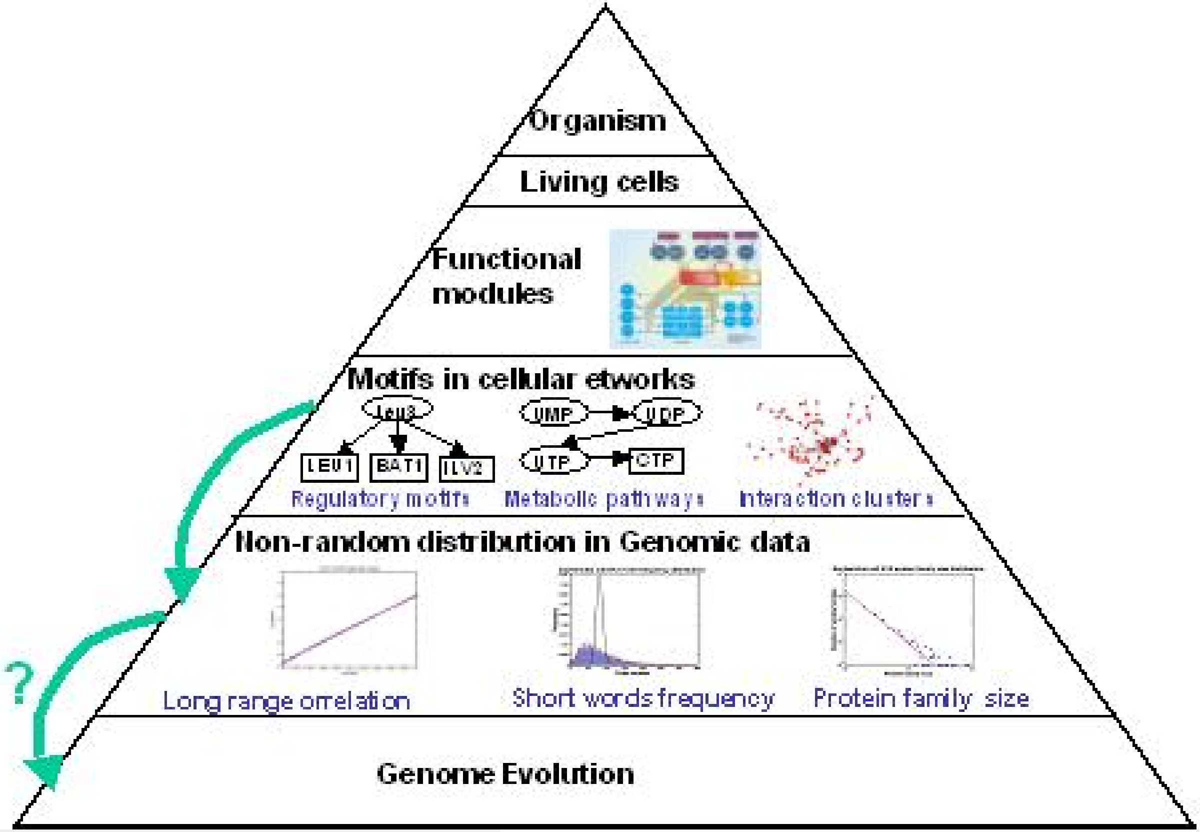
EBD



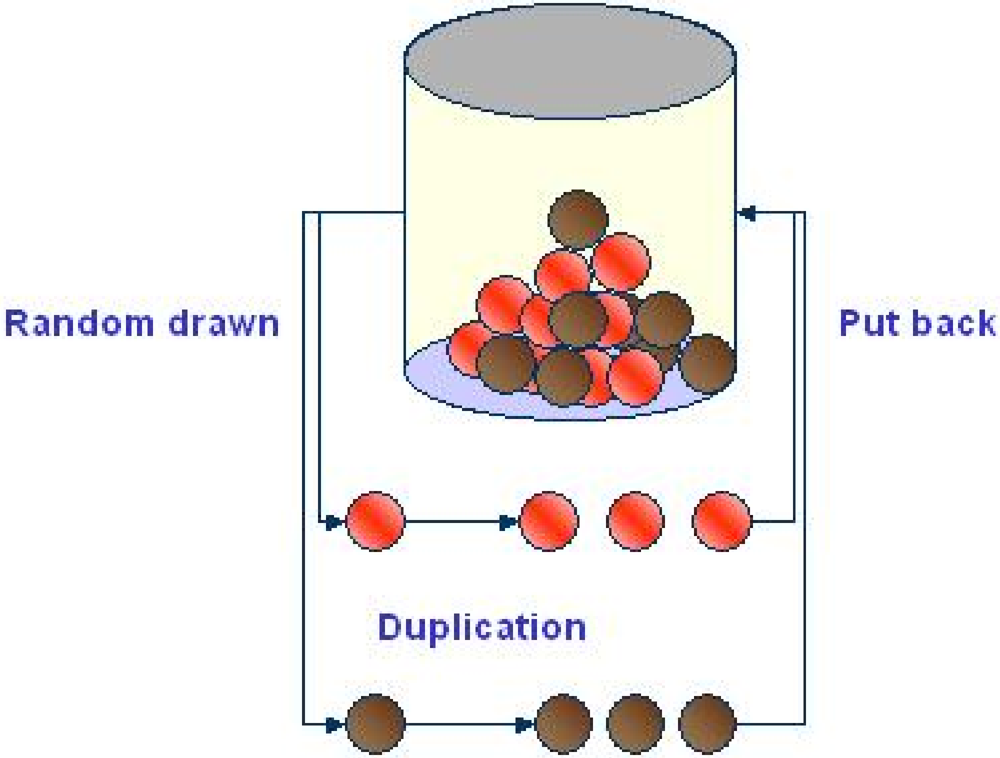
Susumu Ohno

- J.B.S. Haldane (1932):
  - "A redundant duplicate of a gene may acquire divergent mutations and eventually emerge as a new gene."
- Susumu Ohno (1970):
  - "Natural selection merely modified, while redundancy created."

Evolution by Duplication

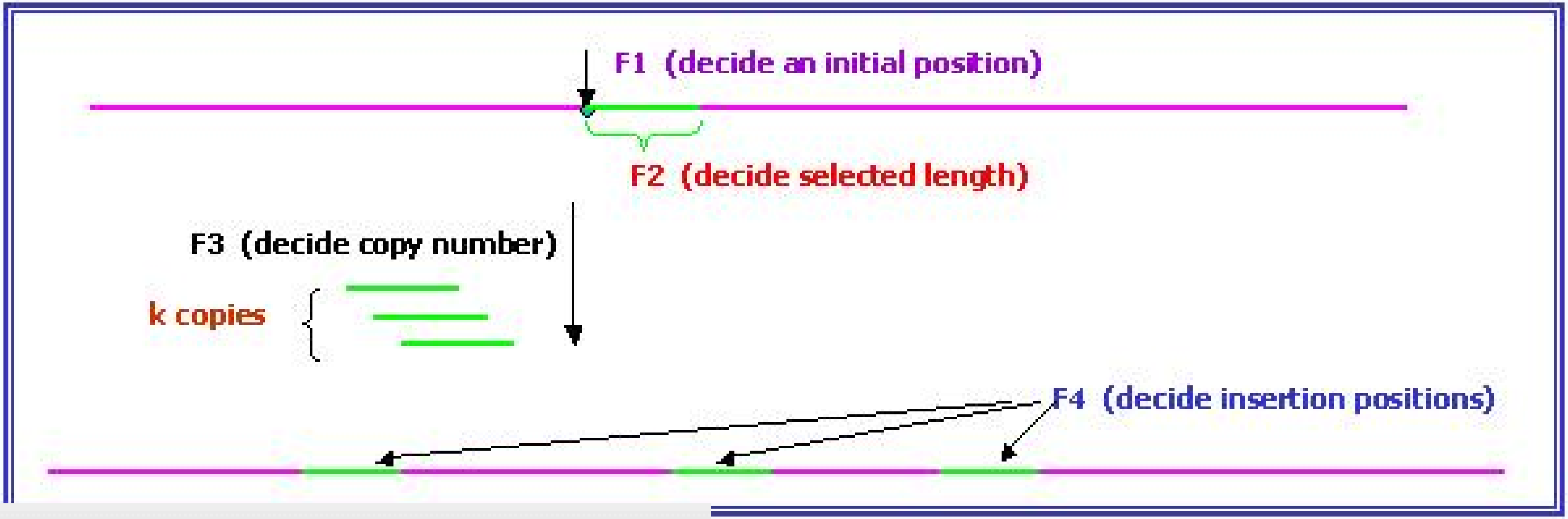


Polya's Urn

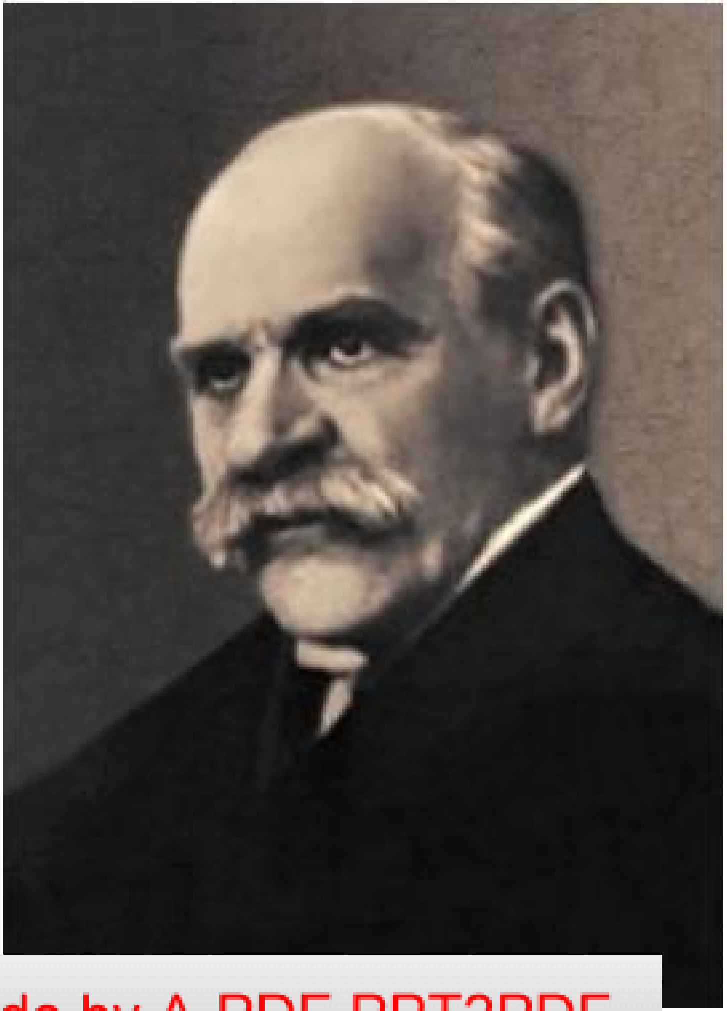


Repetitive Random Eccentric GOD

- Genome Organizing Devices (GOD)
- Polya's Urn Model:
  - F's: functions deciding probability distributions


F1 (decide an initial position)
F2 (decide selected length)
F3 (decide copy number)
k copies
F4 (decide insertion positions)

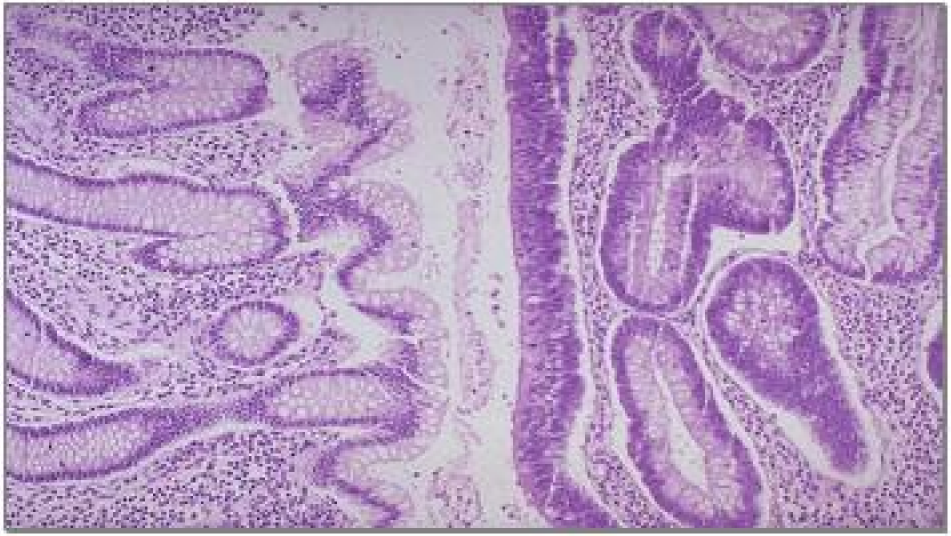
J.B.S. Haldane



- "If I were compelled to give my own appreciation of the evolutionary process…, I should say this: In the first place it is very beautiful. In that beauty, there is an element of tragedy…In an evolutionary line rising from simplicity to complexity, then often falling back to an apparently primitive condition before its end, we perceive an artistic unity …
- "To me at least the beauty of evolution is far more striking than its purpose."
  - J.B.S. Haldane, The Causes of Evolution. 1932.

Human Cancer Genome

Cancer



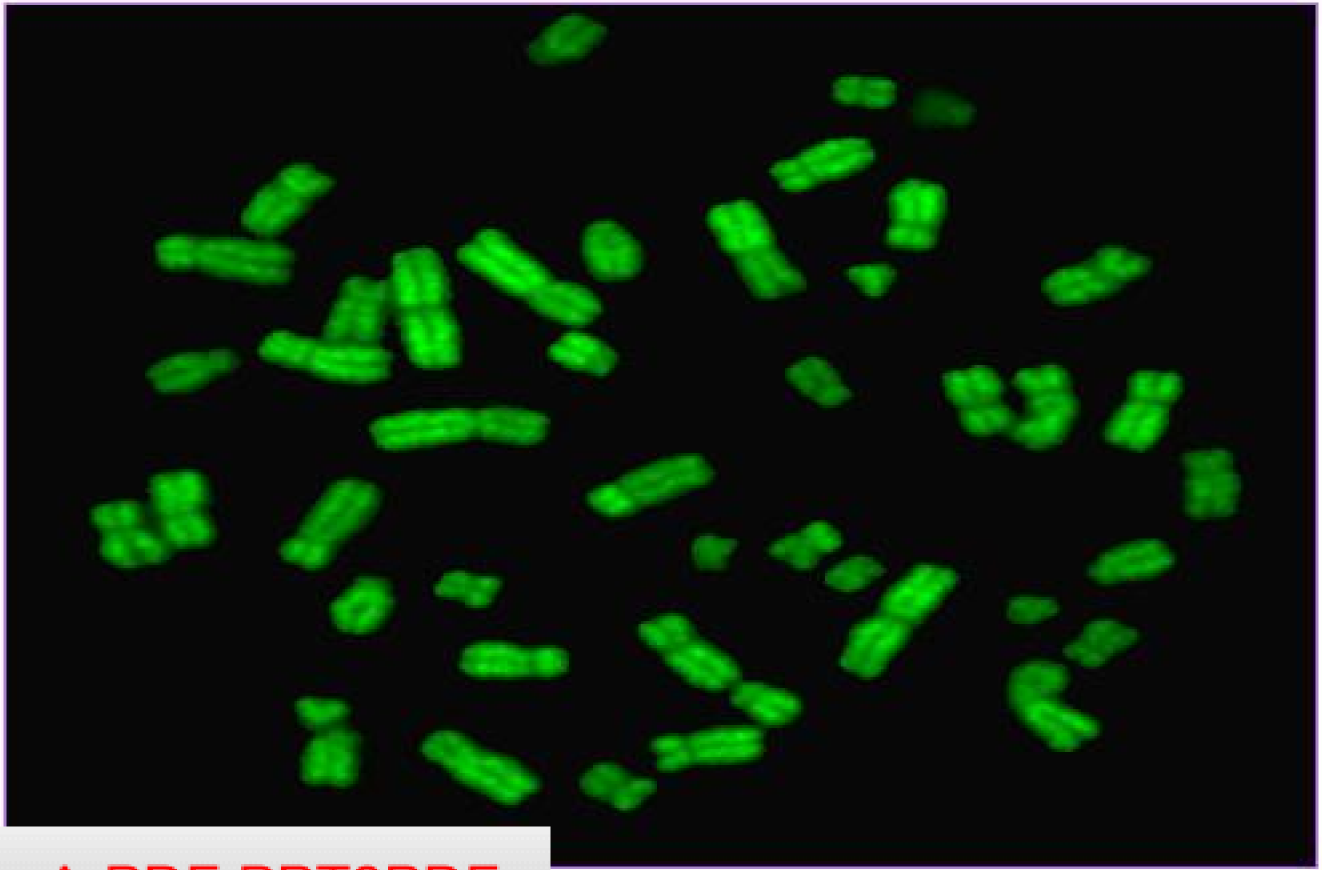
Normal epithelial mucosa

Neoplastic polyp

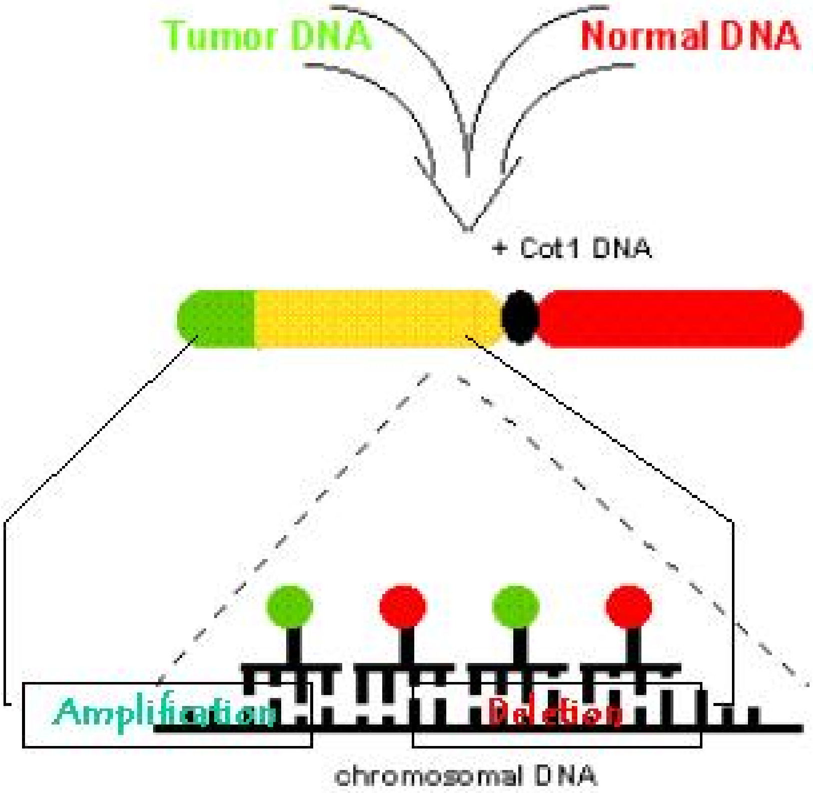
A Challenge

- “At present, description of a recently diagnosed tumor in terms of its underlying genetic lesions remains a distant prospect. Nonetheless, we look ahead 10 or 20 years to the time when the diagnosis of all somatically acquired lesions present in a tumor cell genome will become a routine procedure.”
  - Douglas Hanahan and Robert Weinberg
    - *Cell*, Vol. 100, 57-70, 7 Jan 2000

Karyotyping

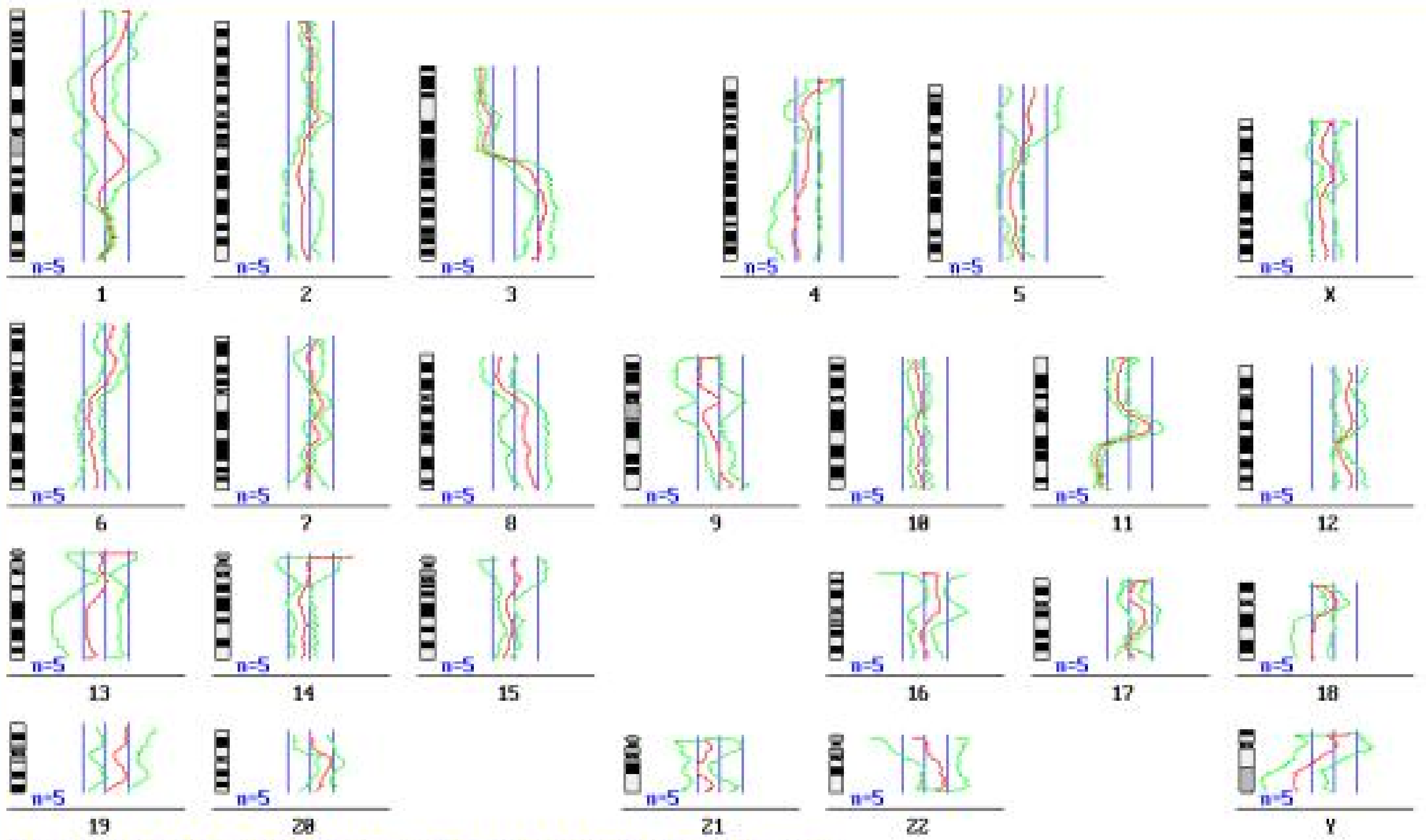


CGH:Comparative Genomic Hybridization.



- 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

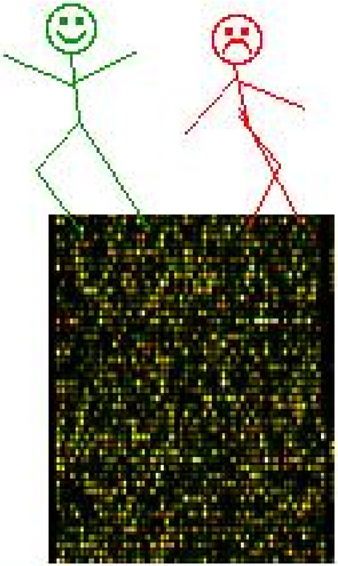
CGH: Comparative Genomic Hybridization.



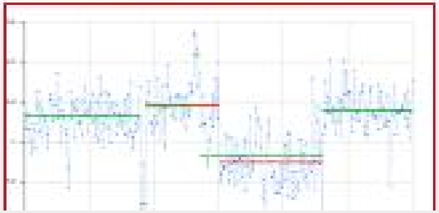
(c) Institute of Pathology, Charite, Humboldt-University Berlin, Germany

Measuring gene copy number differences between complex genomes

Genomic DNA patient samples



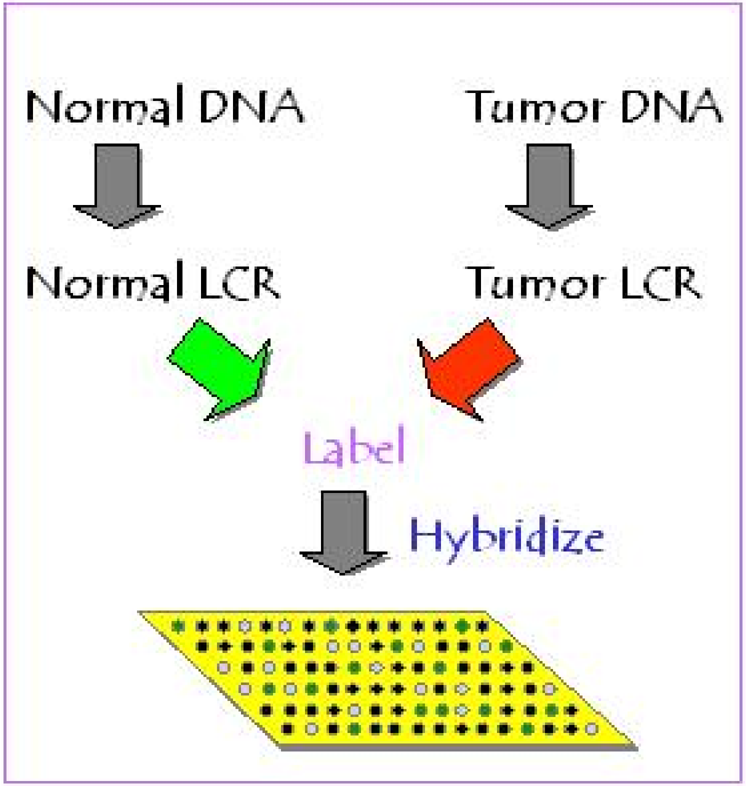
Hybridize to 500,000 features microarray



Statsitical Analysis

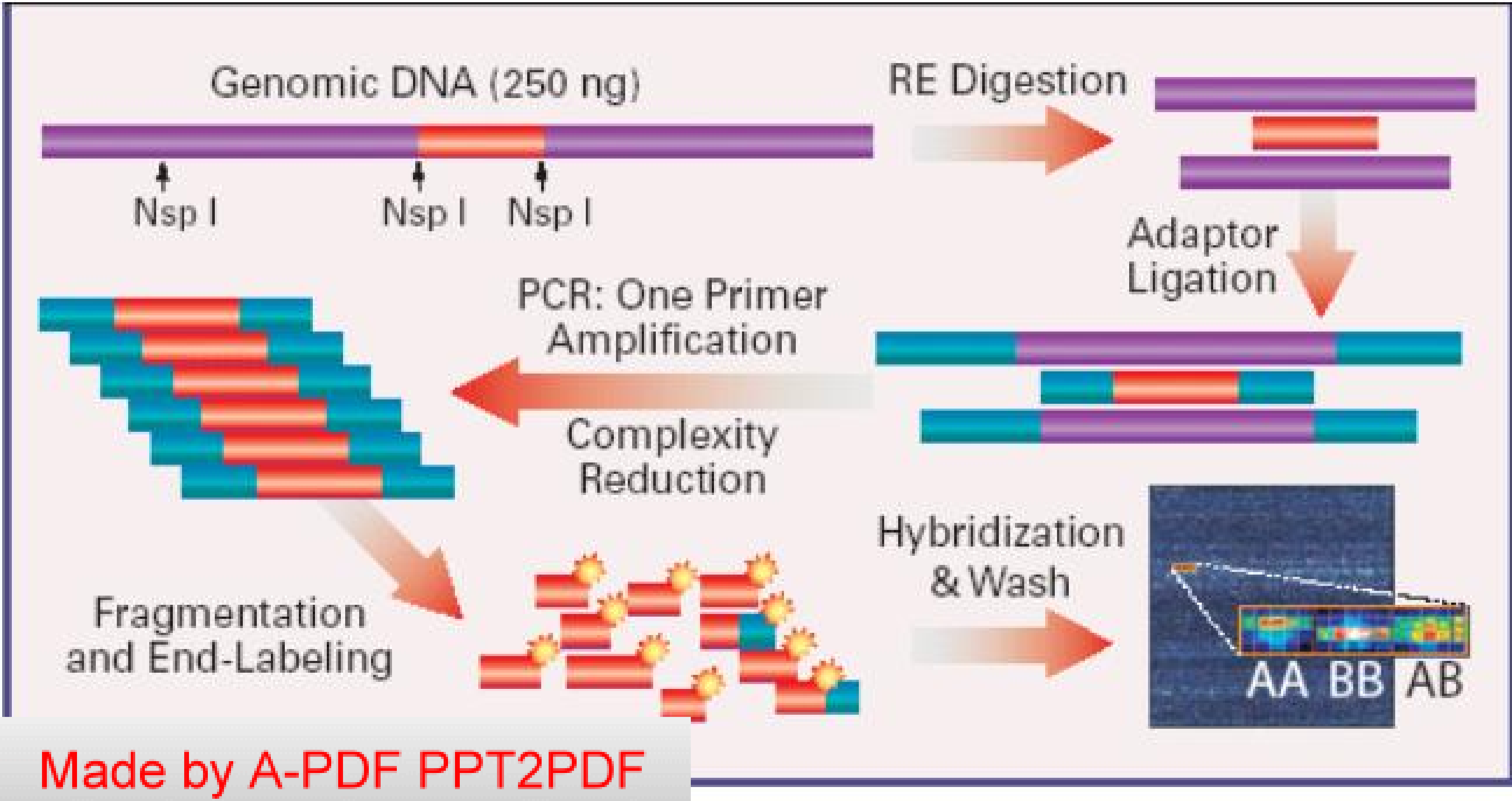
- Compare the genomes of diseased and normal samples
- Error Control:
  - The use of representations augmenting microarrays
  - Representations reproducibly sample the genome thereby reducing its complexity. This increases the signal-to-noise ratio and improves sensitivity
  - Statistical Modeling the sources of Noise
  - Bayesian Analysis

Microarray Analysis of Cancer Genome

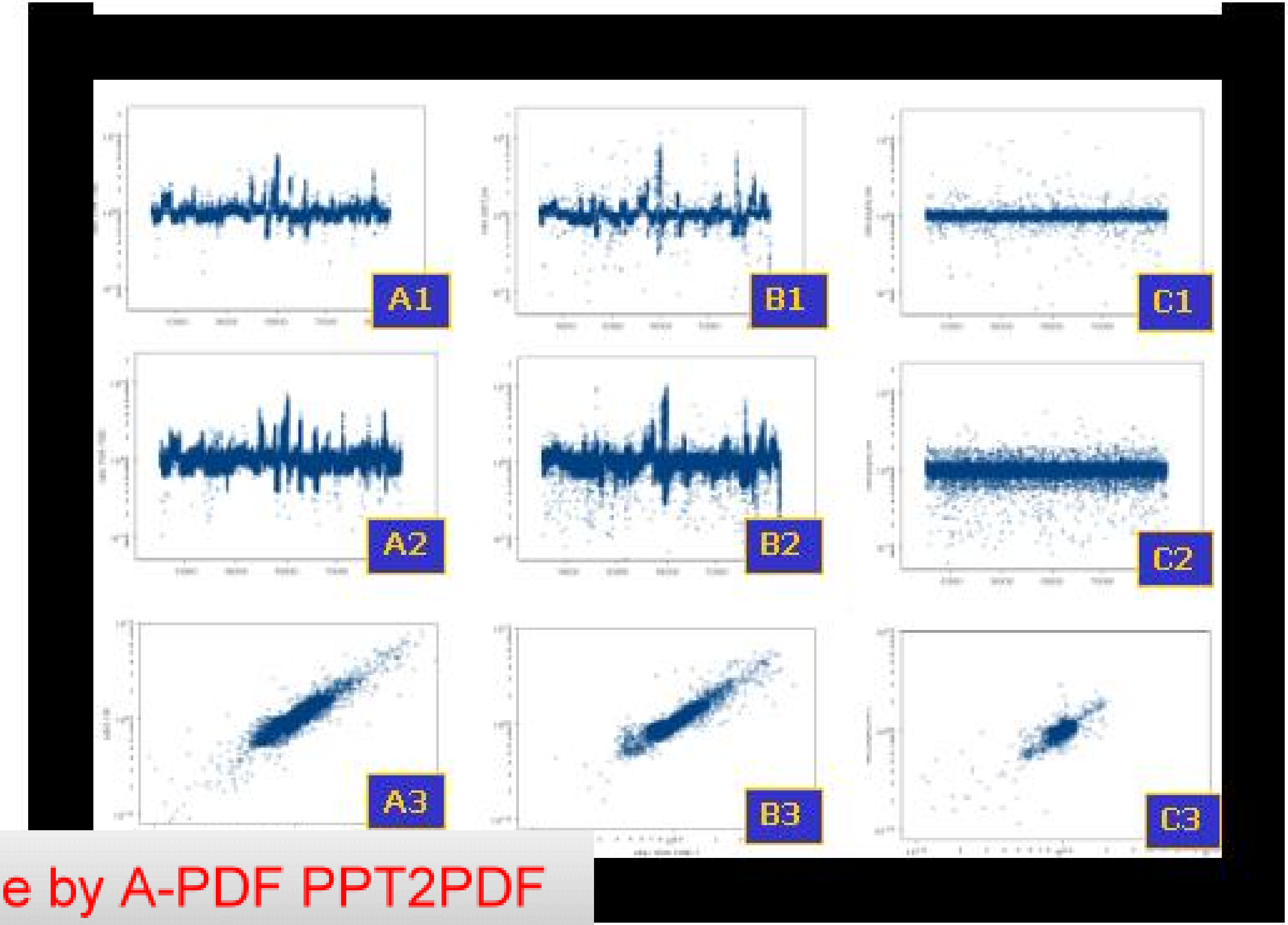


- Representations are reproducible samplings of DNA populations in which the resulting DNA has a new format and reduced complexity.
  - We array probes derived from low complexity representations of the normal genome
  - We measure differences in gene copy number between samples ratiometrically
  - Since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

Minimizing Cross Hybridization (Complexity Reduction)

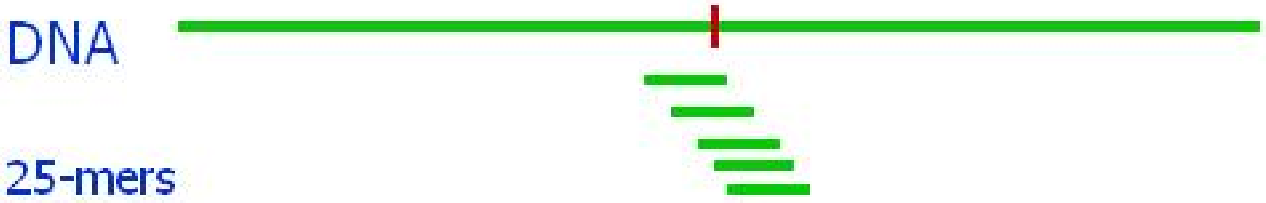


Copy Number Fluctuation



Oligo Arrays: SNP genotyping

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

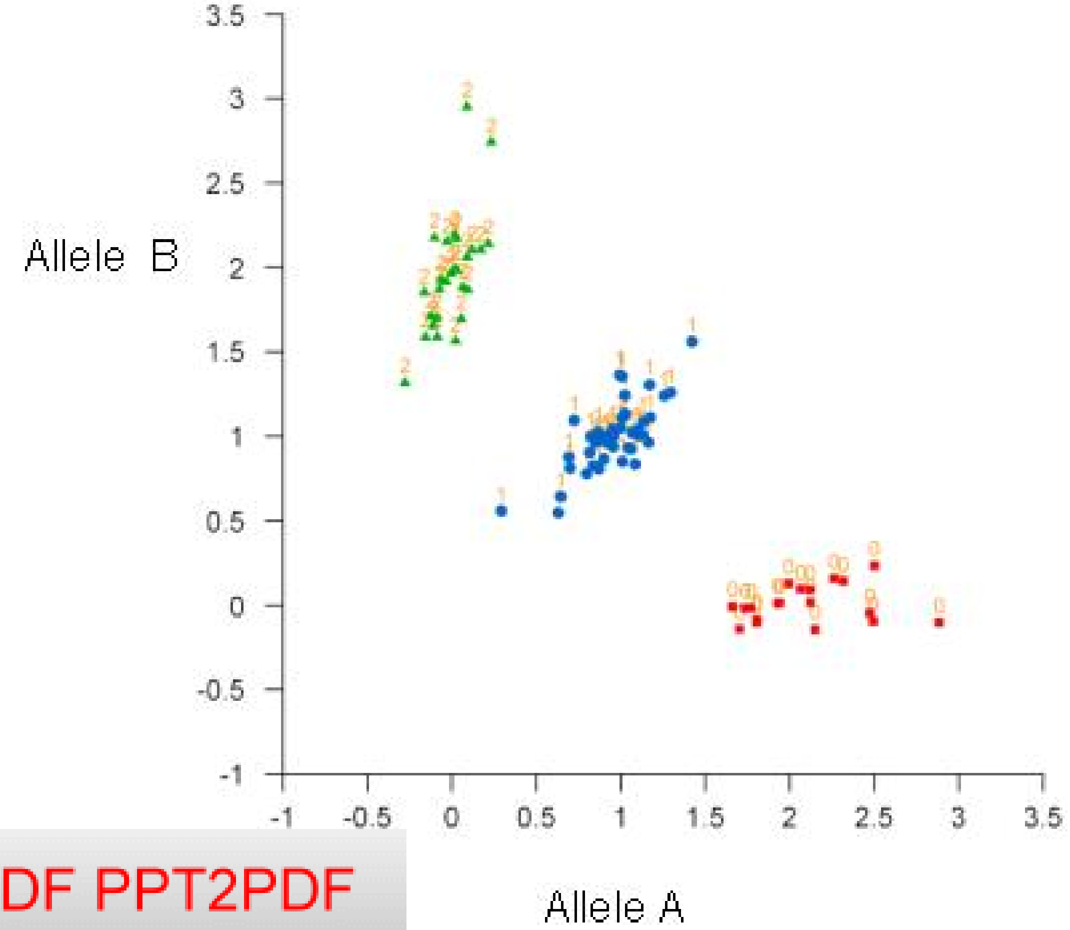


  - Select another 10 25-mers corresponding to SNP Allele B.
  - Problem : Cross Hybridization

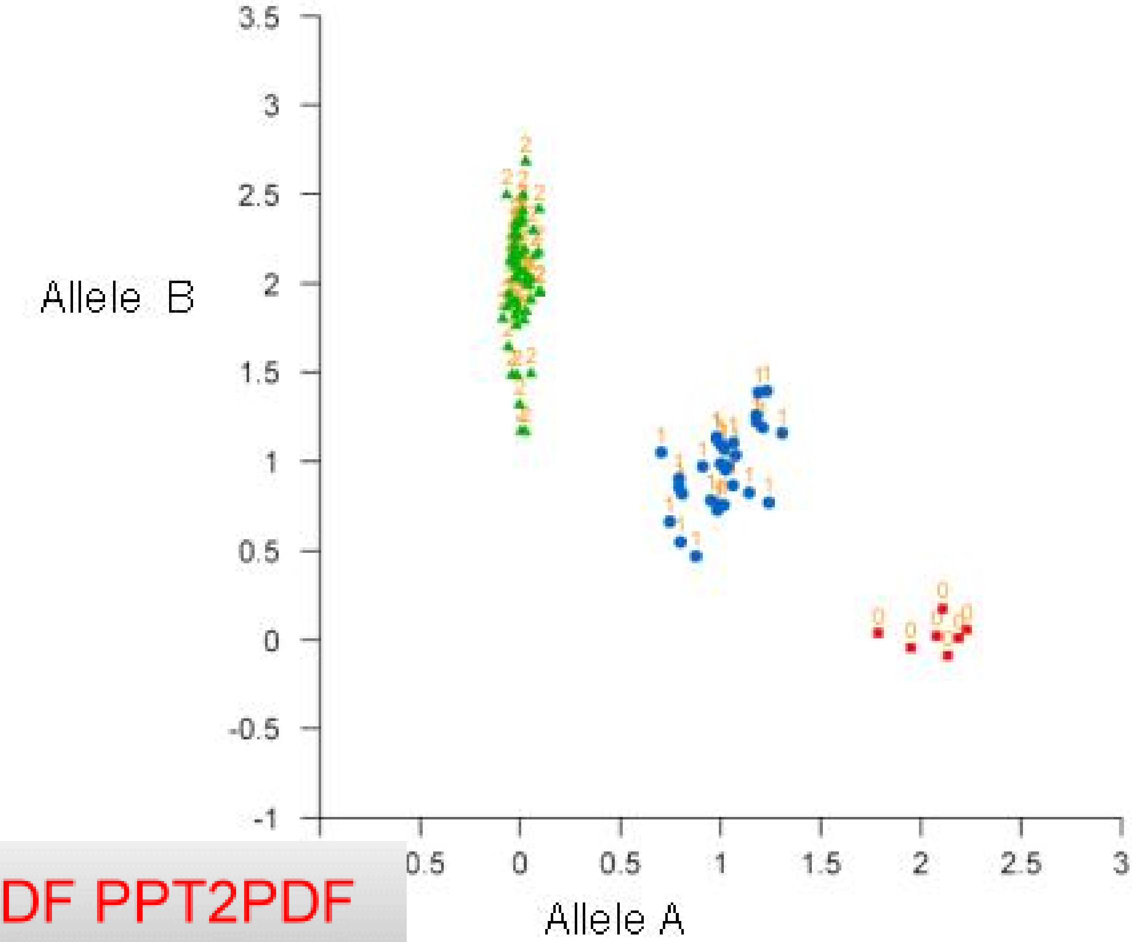
Using SNP arrays to detect Genomic Aberrations

- 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.

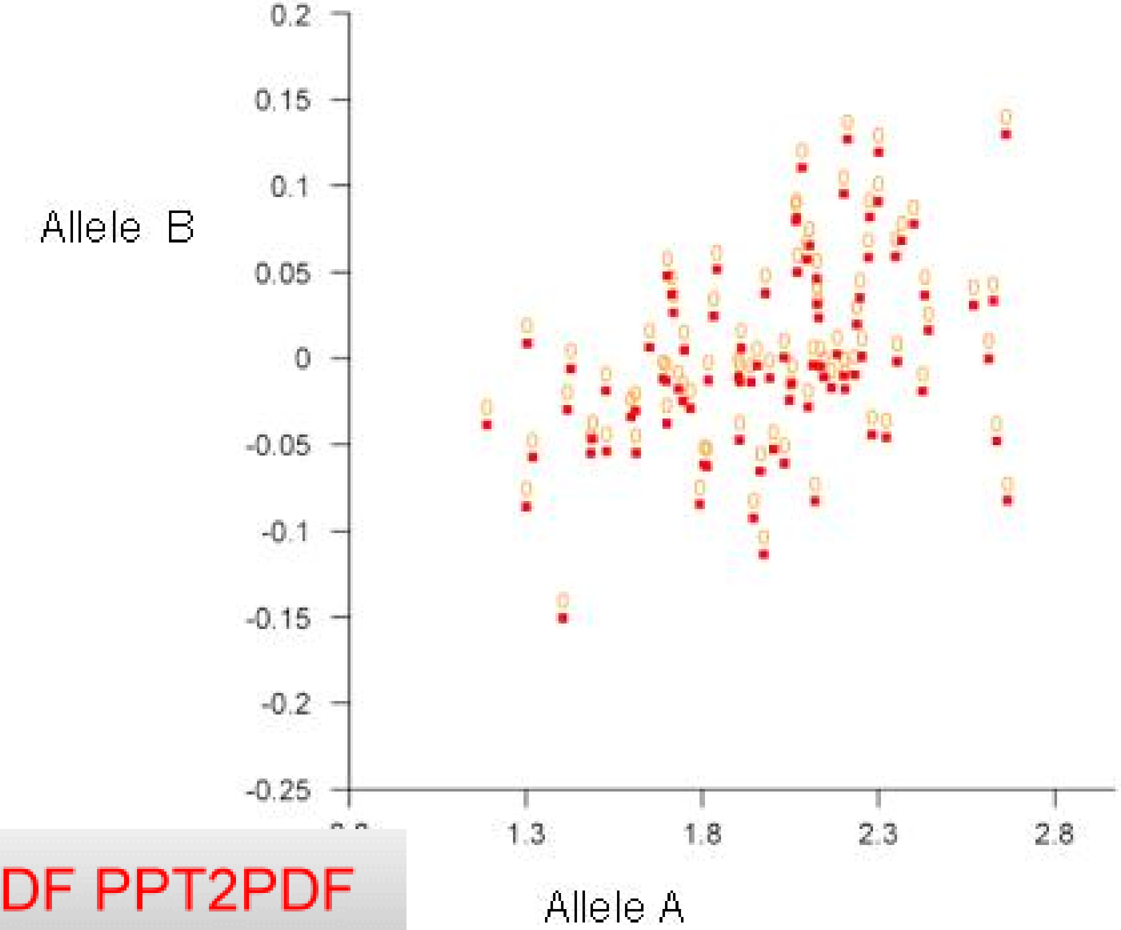
90 humans, 1 SNP (A=0.48)



90 humans, 1 SNP (A=0.24)



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 $p_x \in s_{L,x}$; $p_y \in s_{L,y}$; $p_{x'}, p_{y'} \neg\in G$.
  - After PCR amplification, we have some $K_x \cdot \theta_x$ amount of DNA that is complementary to the probe p_x, etc. K' ($\approx K'_x$) amount of DNA that is additionally approximately complementary to the probe $p_{x'}$.

Normalize using a Generalized RMA

$I' = U - \mu_n$

$- [\alpha\, \sigma_n^{\;2} - \phi_{N(0,1)}(a'/b')/\Phi_{N(0,1)}(a'/b')]$

$\times\{(1 + \beta'\, B_{\sigma_n}/\Phi_{N(0,1)}(a'/b')\}^{-1}$

$+ [b_{\sigma_n}/B_{\sigma_n}]\ )]$

$\times\{(1 + \Phi_{N(0,1)}(a'/b')/(\beta'\, B_{\sigma_n})\}^{-1},$

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

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[$\mu1 + \varepsilon\ \sigma1$], a multiplicative logNormal noise, [$\mu2 + \varepsilon\ \sigma2$] an additive Gaussian noise, and $\phi'_x = K_x\ \phi_x$ an amplified affinity.

- A more general model:

$$I_x = [\theta_x + K'/K_x] \ \phi'_x \ e^{\mu1 + \varepsilon\ \sigma1} + \mu2 + \varepsilon\ \sigma2$$

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

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

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 :

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

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 :

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

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 approprately.

Outlier Model

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

where

$$P_2(PM_{ik}) = \text{Uniform Distributi on}$$

$$\text{w}_1 = \text{Prior probabilit y that data is NOT outlier.}$$

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 :

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

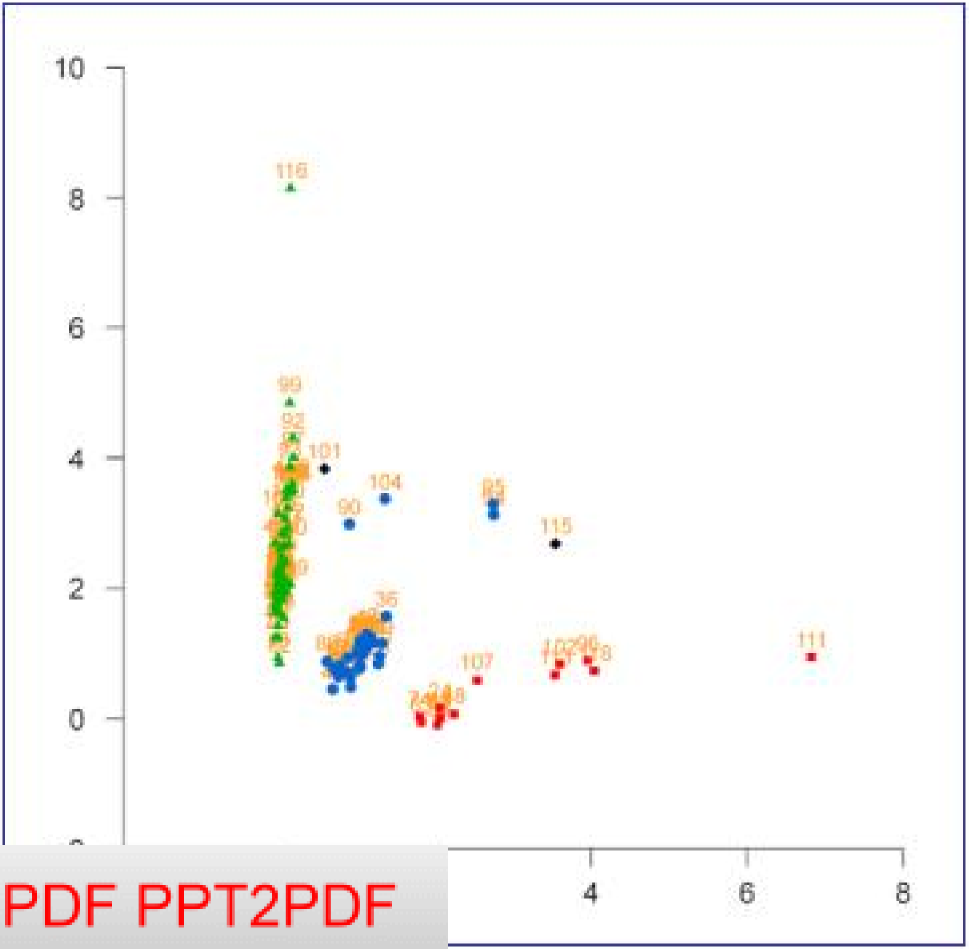
The correct scaling factors C_j, D_j cannot be inferred from the data model.

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

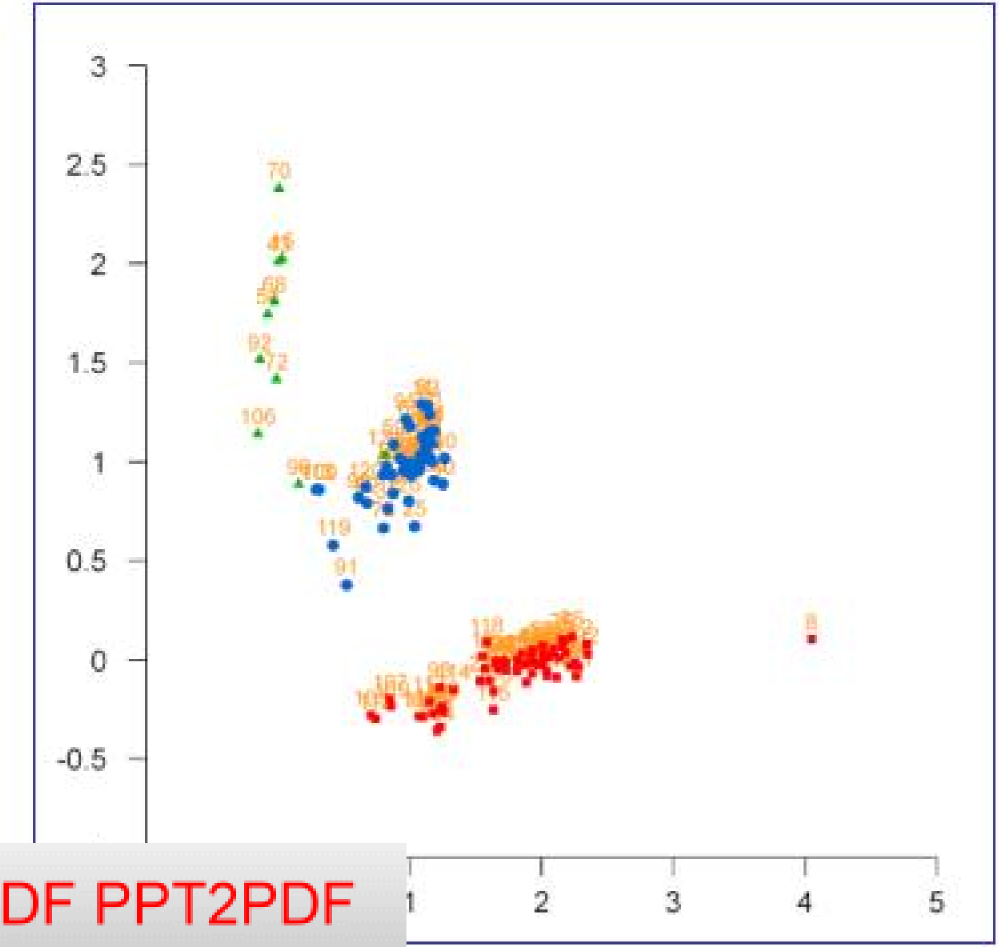
Segmentation to reduce noise

- 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.

Allelic Frequencies: Cancer & Normal



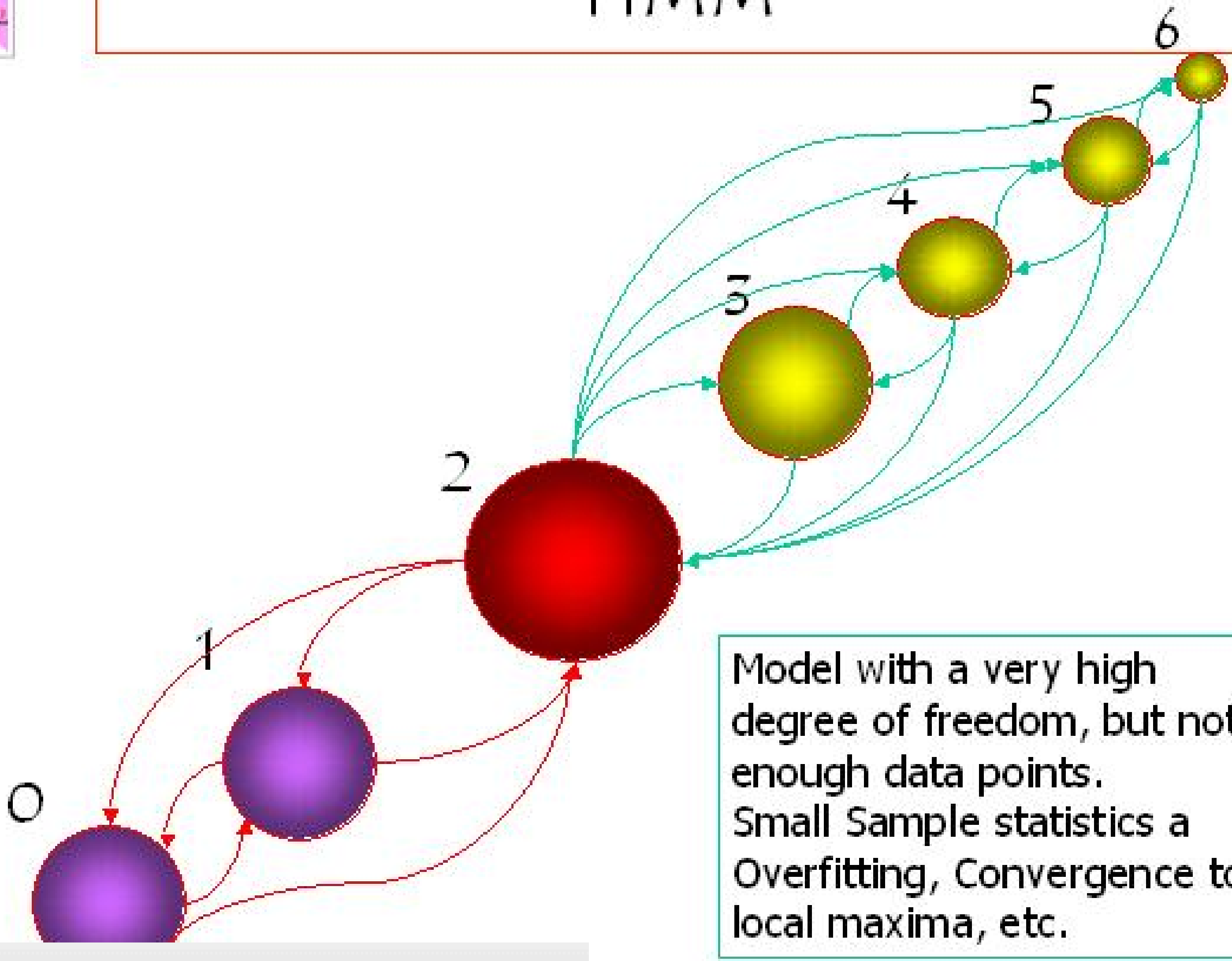
Allelic Frequencies: Cancer & Normal



Segmentation & Break-Point Detection

- 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

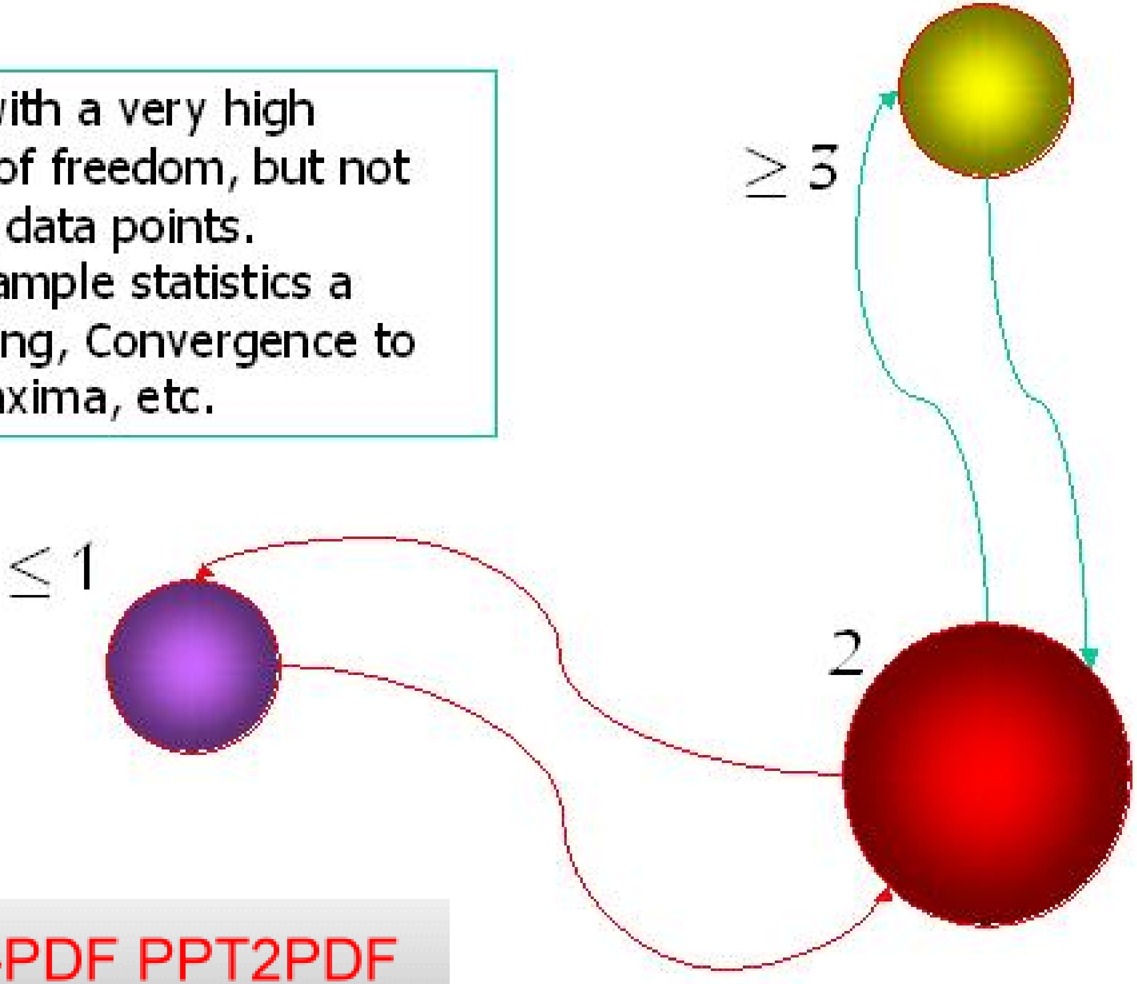
HMM



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

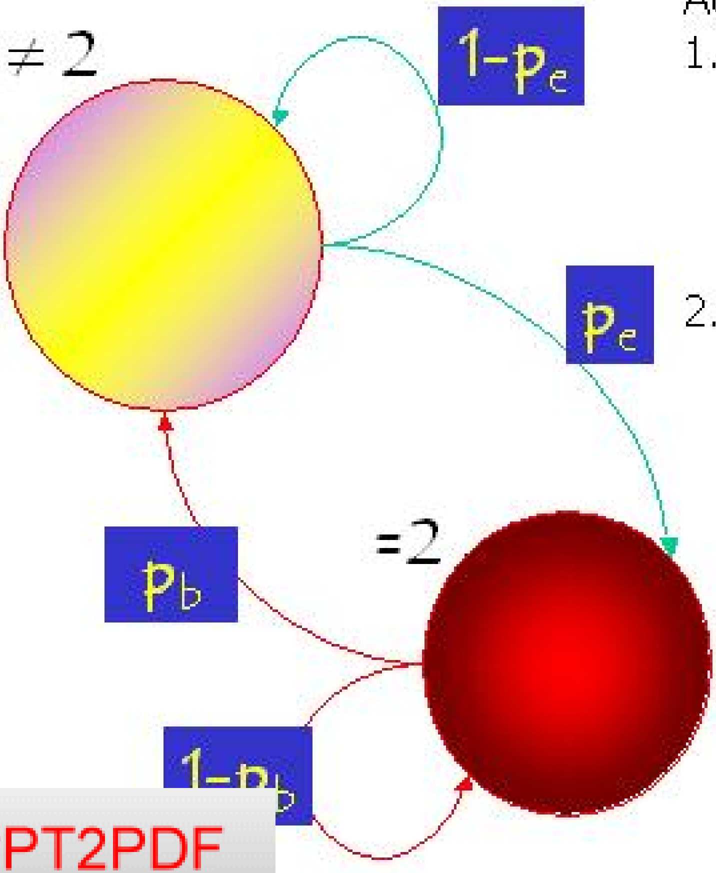
HMM, finally...

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

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:

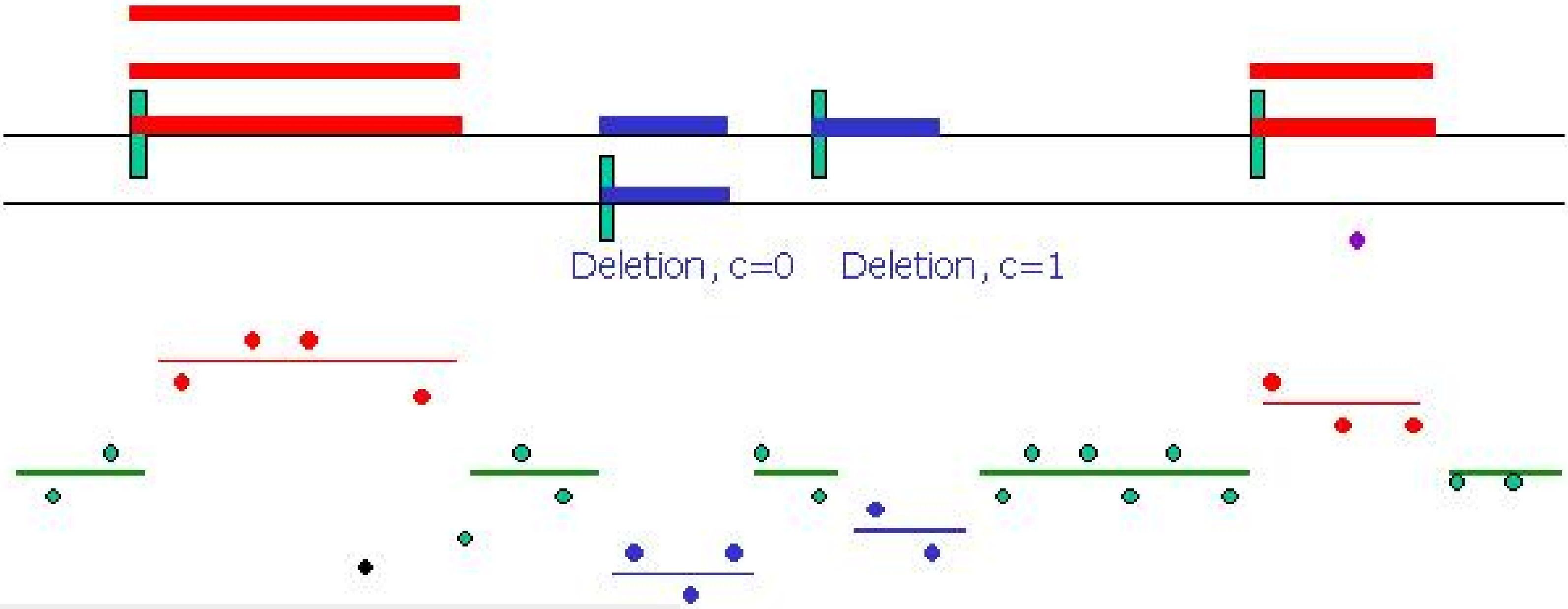
1. Small Number of parameters. Can be optimized by MAP estimator. (EM has difficulties).
2. Easy to model deviation from Markvian properties (e.g., polymorphisms, power-law, Polya's urn like process, local properties of chromosomes, etc.)

Generative Model

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

Amplification, c=4

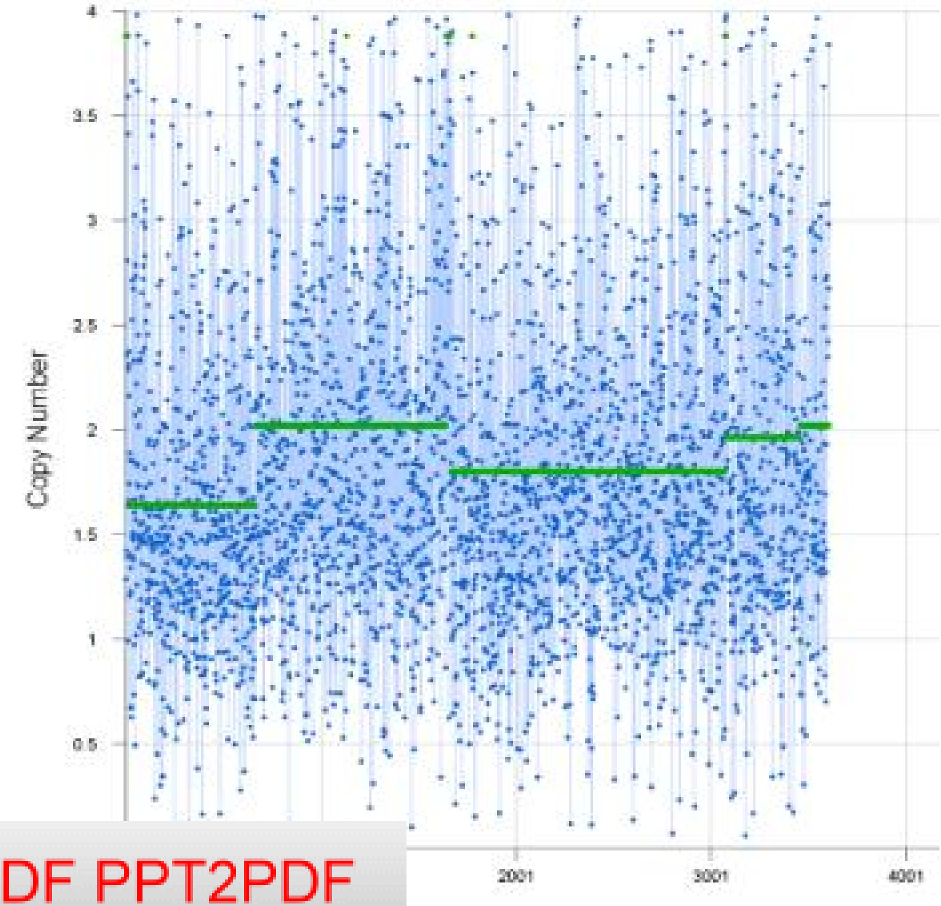
Amplification, c=3



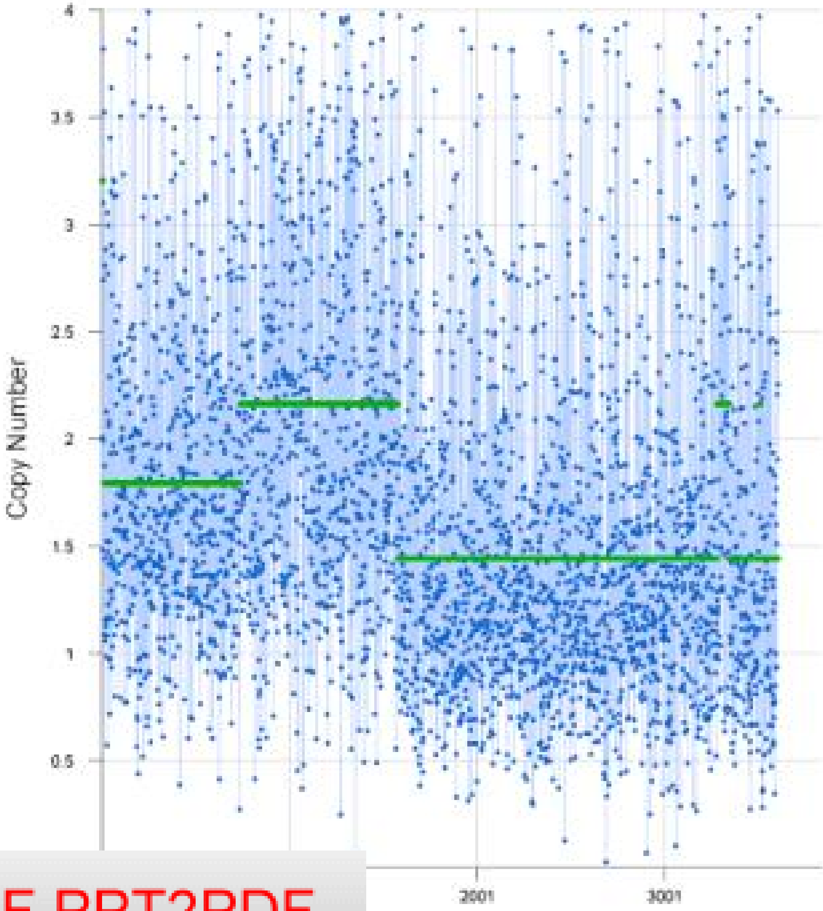
Deletion, c=0

Deletion, c=1

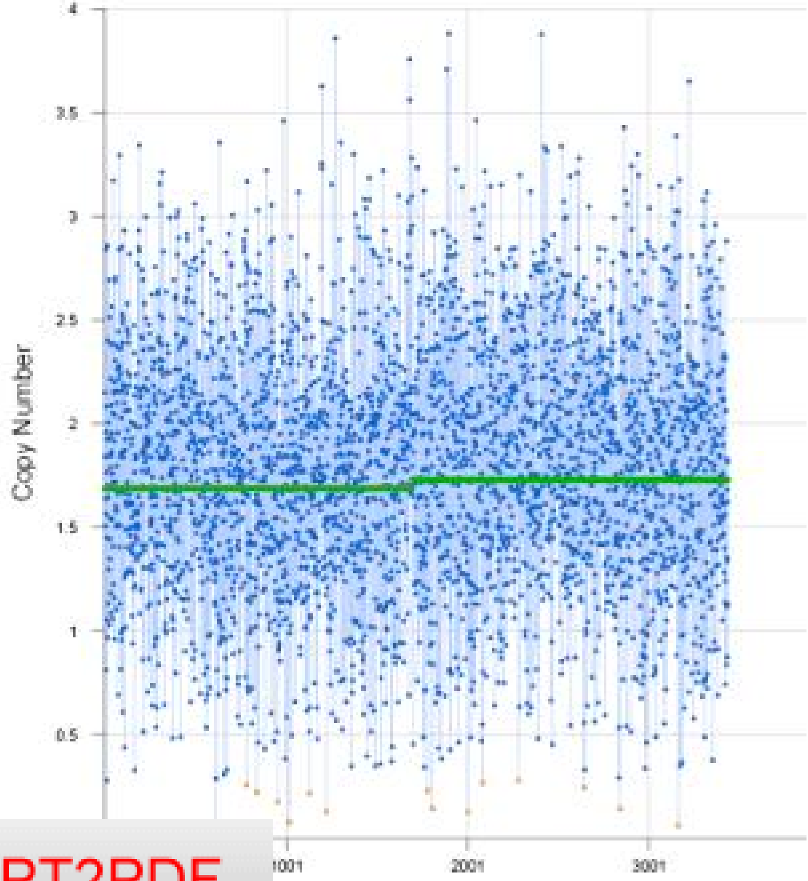
Segmentation Normal χ_7



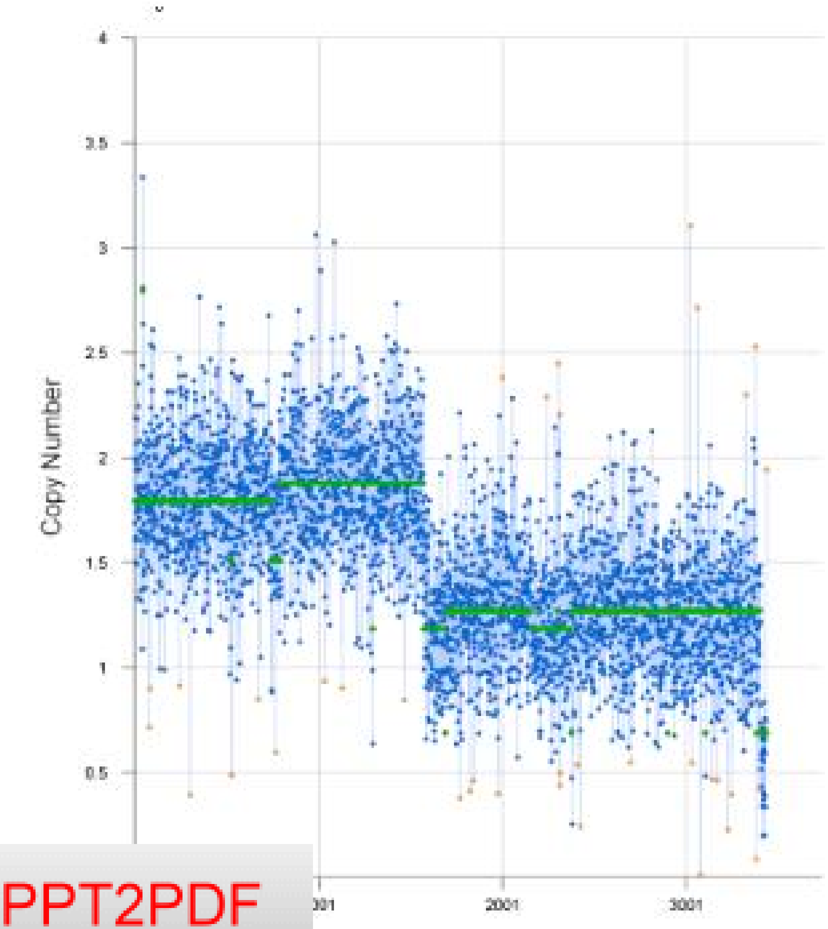
Segmentation Cancer χ_7



Corrected Segmentation Normal χ_7



Corrected Segmentation Cancer χ_7



Likelihood Function

- The likelihood function for first n probes:
- $L(\langle i_1, \mu_1, \ldots, i_k, \mu_k \rangle)$

$$= \mathrm{Exp}(-p_b\, n)\, (p_b\, n)^k$$

$$\times (2\, \pi\, \sigma^2)^{(-n/2)} \prod_{i=1}^{n} \mathrm{Exp}[-(v_i - \mu_j)^2/2\sigma^2]$$

$$\times\, p_e^{(\#global)}(1-p_e)^{(\#local)}$$

  - Where $i_k = n$ and i belongs to the jth interval.
  - Maximum A Posteriori algorithm (implemented as a Dynamic Programming Solution) optimizes L to get the best segmentation
- $L(\langle i^*_1, \mu^*_1, \ldots, i^*_k, \mu^*_k \rangle)$

Dynamic Programming Algorithm

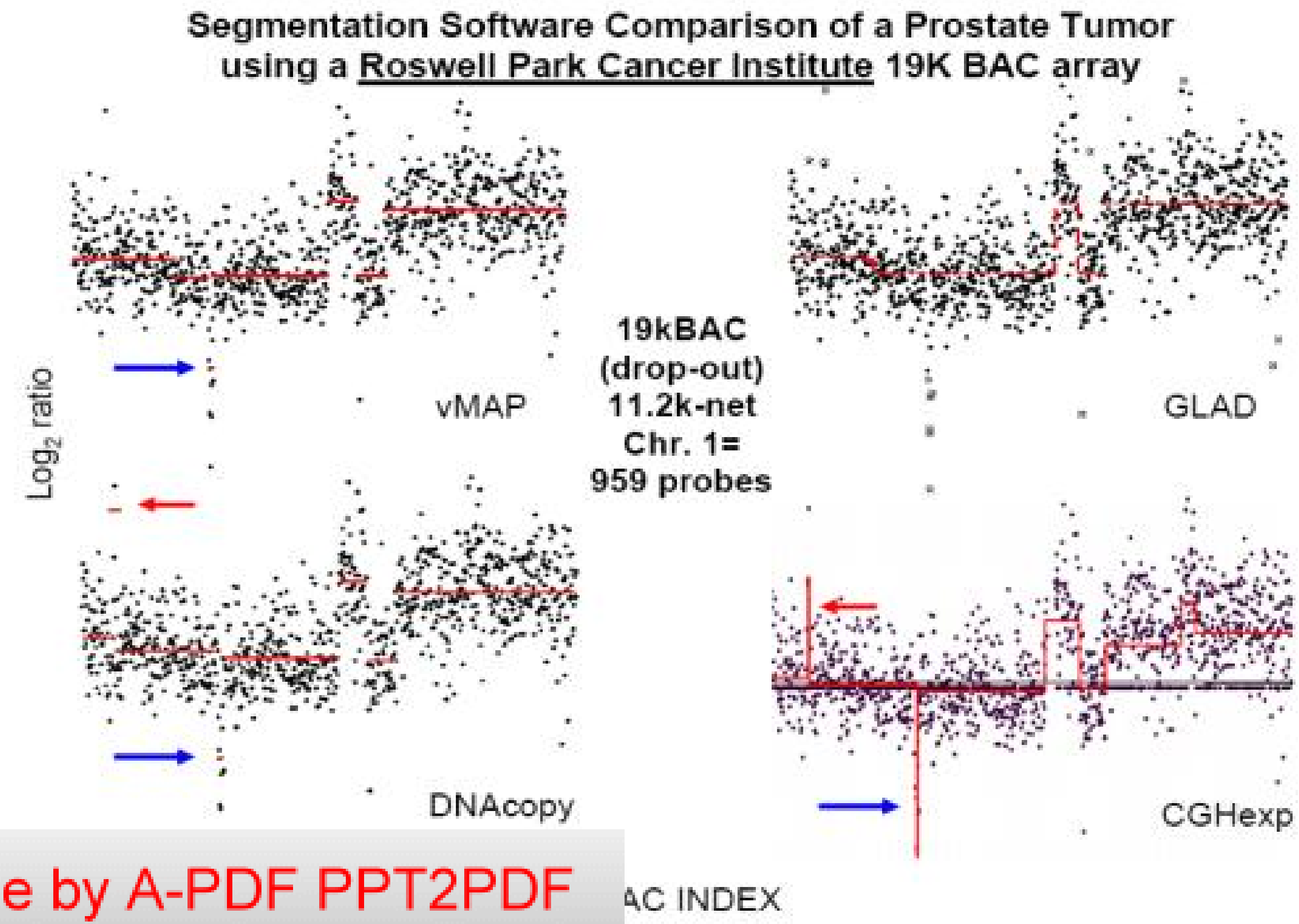
- Generalizes Viterbi and Extends.
- Uses the optimal parameters for the generative model:
- Adds a new interval to the end:
- $\langle i_1, \mu_1, \ldots, i_k, \mu_k \rangle \circ \langle i_{k+1}, \mu_{k+1} \rangle = \langle i_1, \mu_1, \ldots, i_k, \mu_k, i_{k+1}, \mu_{k+1} \rangle$
- Incremental computation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

$$
\begin{aligned}
& - \mathrm{Log}\ L(\langle i_1, \mu_1, \ldots, i_k, \mu_k, i_{k+1}, \mu_{k+1} \rangle) \\
& = -\mathrm{Log}\ L(\langle i_1, \mu_1, \ldots, i_k, \mu_k \rangle) \\
& + \text{new-res.}/2\sigma^2 - \mathrm{Log}(p_b n) + (i_{k+1} - i_k)\ \mathrm{Log}\ (2\pi\sigma^2) \\
& - (i_{k+1} - i_k)\ [\mathbb{I}_{global}\ \mathrm{Log}\ p_e + \mathbb{I}_{local}\ \mathrm{Log}(1 - p_e)]
\end{aligned}
$$

Segmentation Analysis

NAME	Algorithm	Front-end	Published
VMAP	Bayes / t-statistic	WEB	Daruwala et al. PNAS Sci U S A, 2004
DNAcopy	Circular binary segmentation	R, Java	Olshen, AB et al. Biostatistics, 2004
GLAD	Adaptive Weights Smoothing	R, Java	Lingjaerde, OC et al. Bioinformatics, 2005
CGH-explorer	Analysis of Copy Errors	Java	Hupe, P et al. Bioinformatics, 2004
Recent reviews:	Willenbrock & Fridlyand, Bioinformatics, 2005; Weil et al. Bioinformatics, 2005		

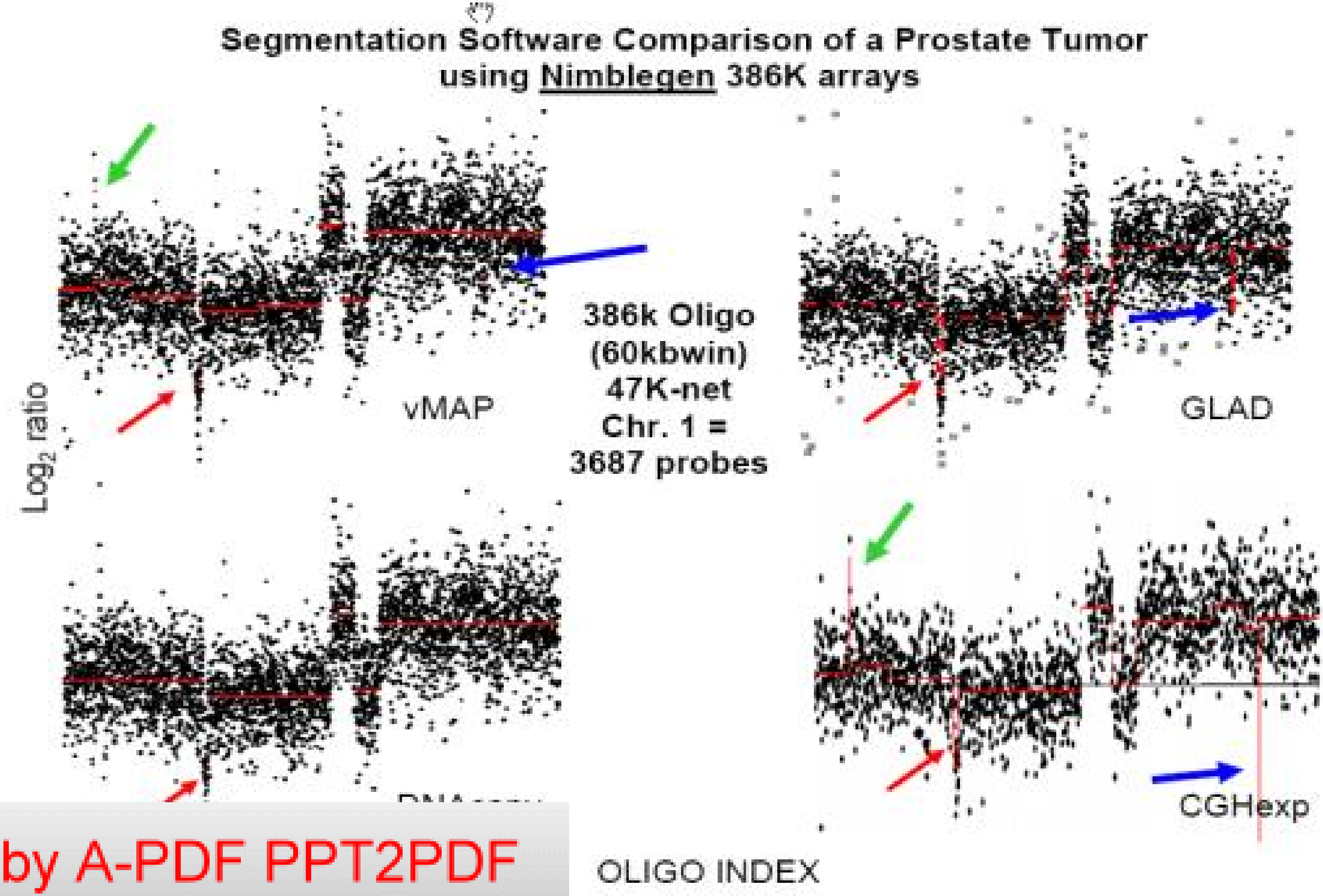
Comparative Analysis: BAC Array



Segmentation Software Comparison of a Prostate Tumor using a Roswell Park Cancer Institute 19K BAC array

19kBAC (drop-out) 11.2k-net Chr. 1= 959 probes

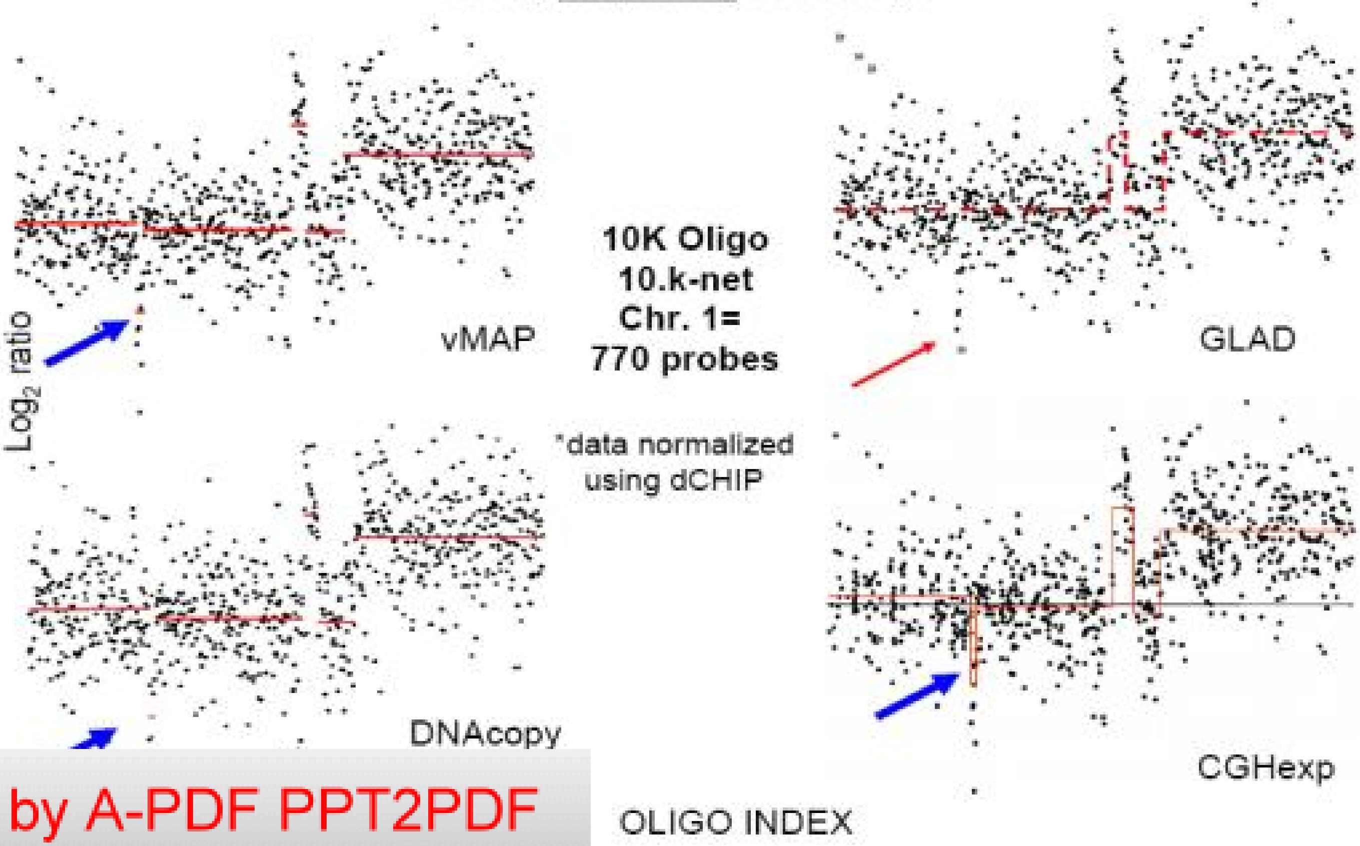
Comparative Analysis: Nimblegen



Segmentation Software Comparison of a Prostate Tumor using Nimblegen 386K arrays

Comparative Analysis: Affy 10K

Segmentation Software Comparison of a Prostate Tumor using Affymetrix 10K Arrays



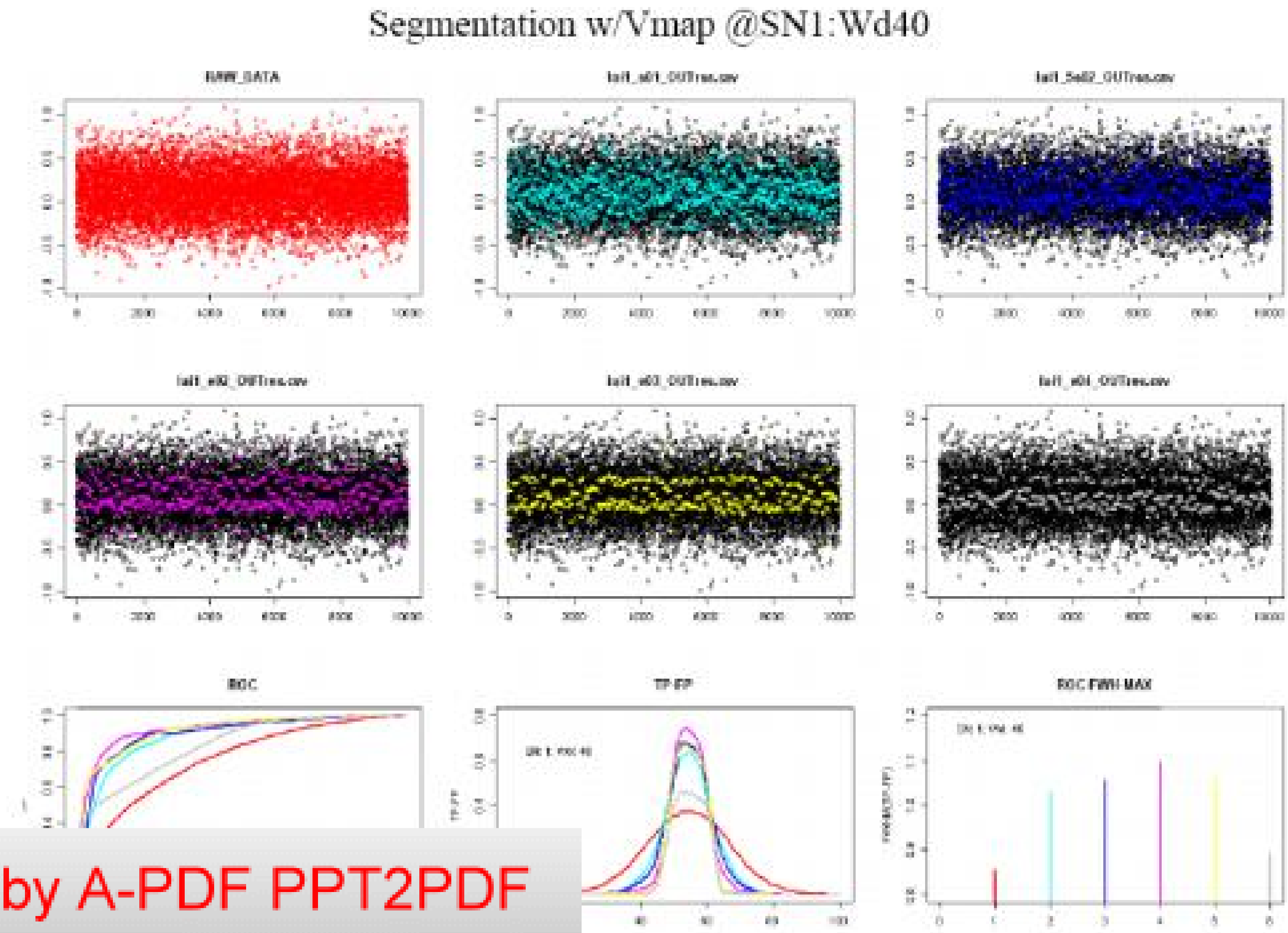
10K Oligo
10.k-net
Chr. 1=
770 probes

*data normalized using dCHIP

Simulated Data

- Array CGH simulations and an "ROC analysis"
  - Using the same scheme as Lai et al.
    - Weil R. Lai, Mark D. Johnson, Raju Kucherlapati, and Peter J. Park (2005), "Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data," *Bioinformatics*, **21**(19): 3763-3770.
- Segmented by Vmap and DNAcopy
    - Vmap algorithm was tested at 11 segmentation Pvalues of: 0.1, $5\ 10^{-2}$, 10^{-2}, 10^{-3}, 10^{-4}, …, 10^{-10}.
    - DNAcopy algorithm was tested at 9 segmentation alpha values of: .9, .5, .1, 10^{-2}, 10^{-3}, 10^{-4}, …, 10^{-7}.
- Analysis by Alex Pearlman et al. (2006)

VMAP



Segmentation w/Vmap @SN1:Wd40

RAW_DATA

lail_a01_OUTres.csv

lail_5a02_OUTres.csv

lail_a02_OUTres.csv

lail_a03_OUTres.csv

lail_a04_OUTres.csv

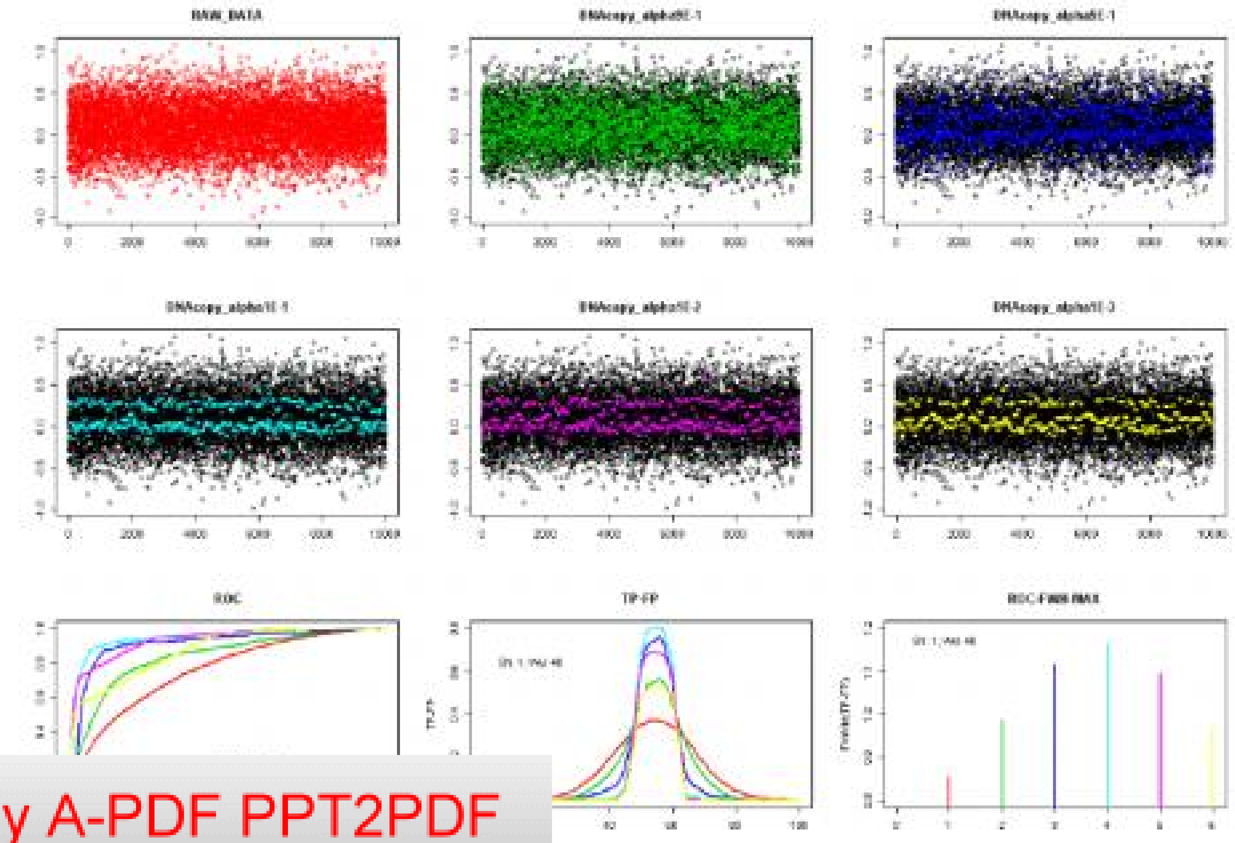
ROC

TP-FP

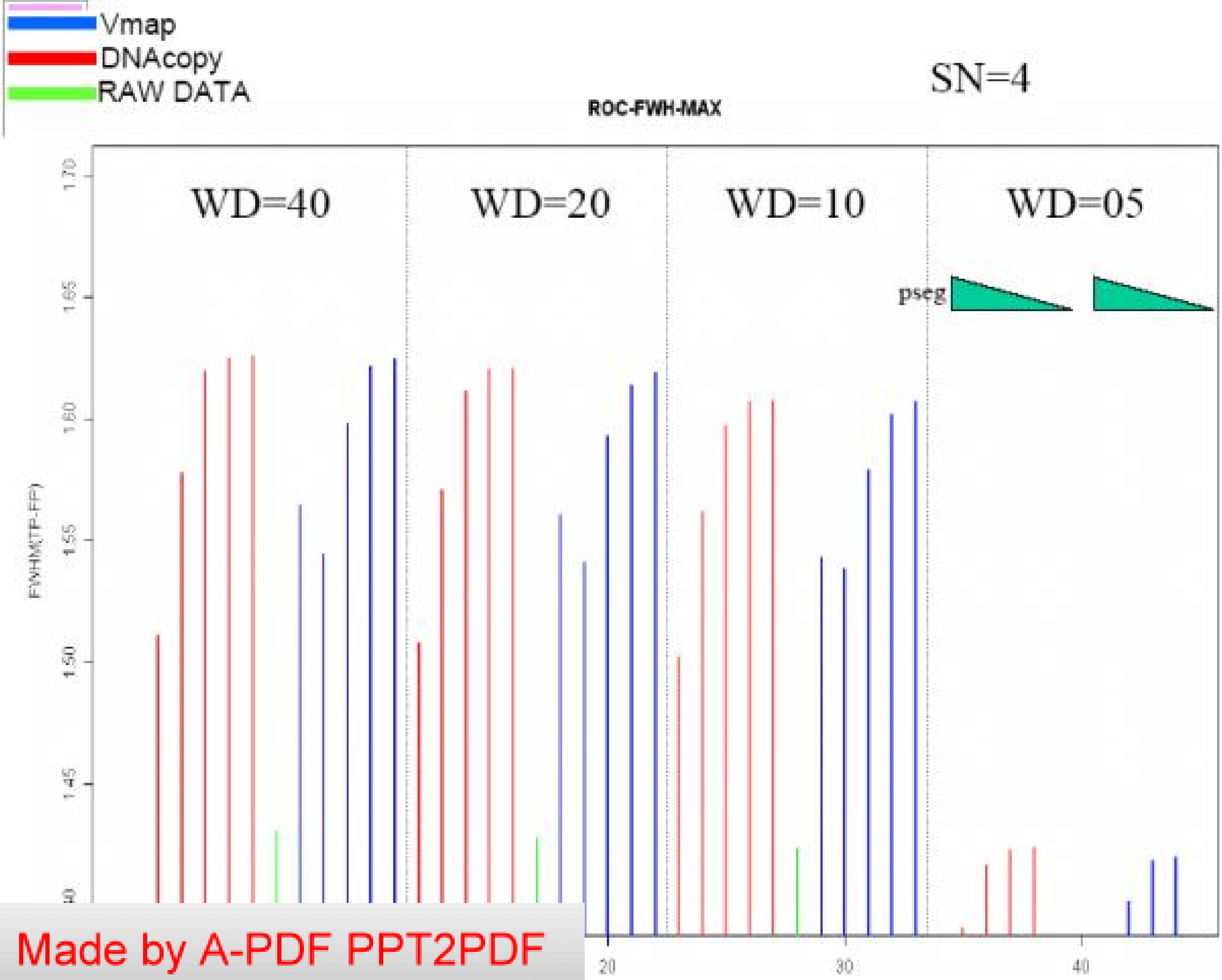
ROC FWH MAX

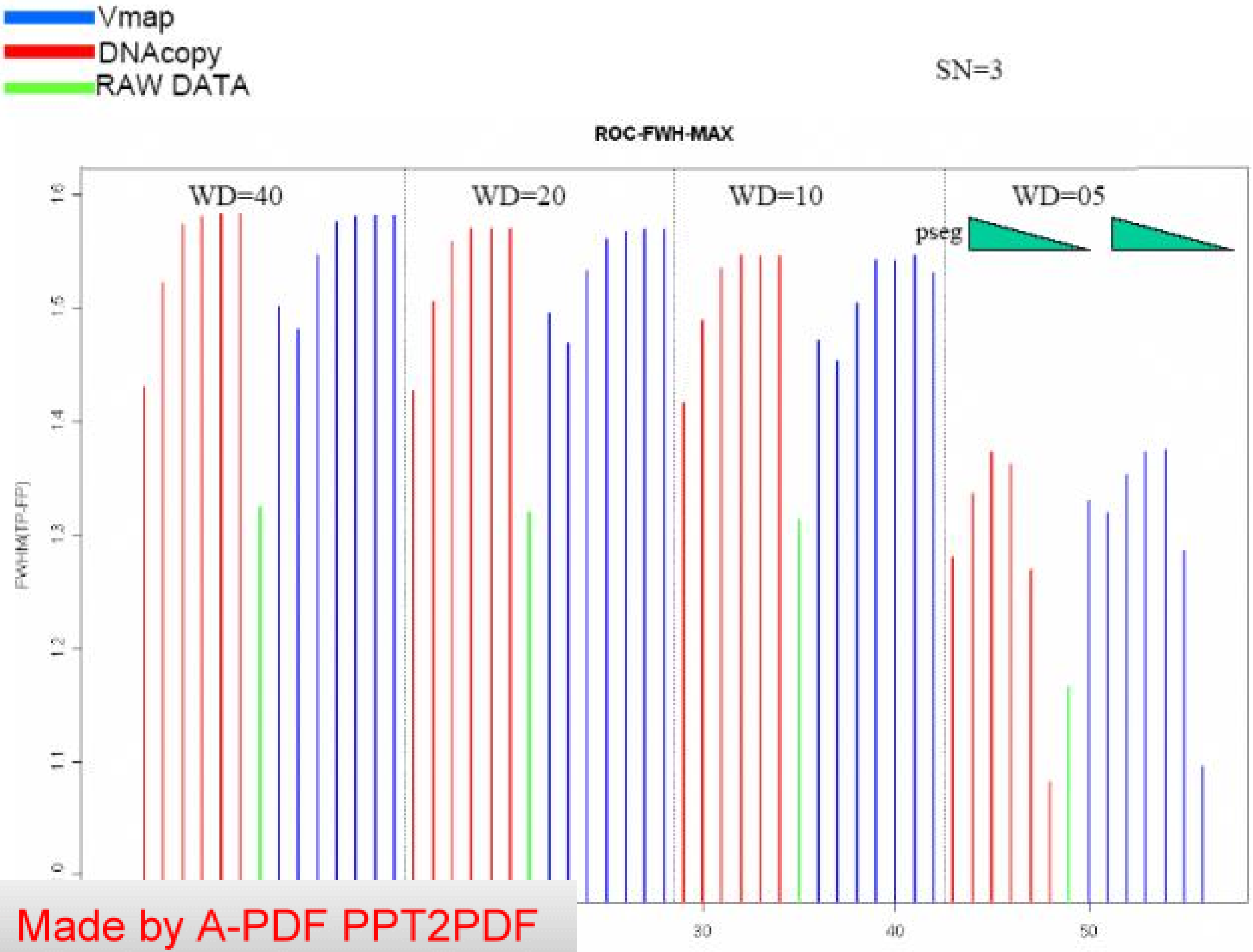
DNACopy

Segmentation w/DNAcopy @SN1:Wd40

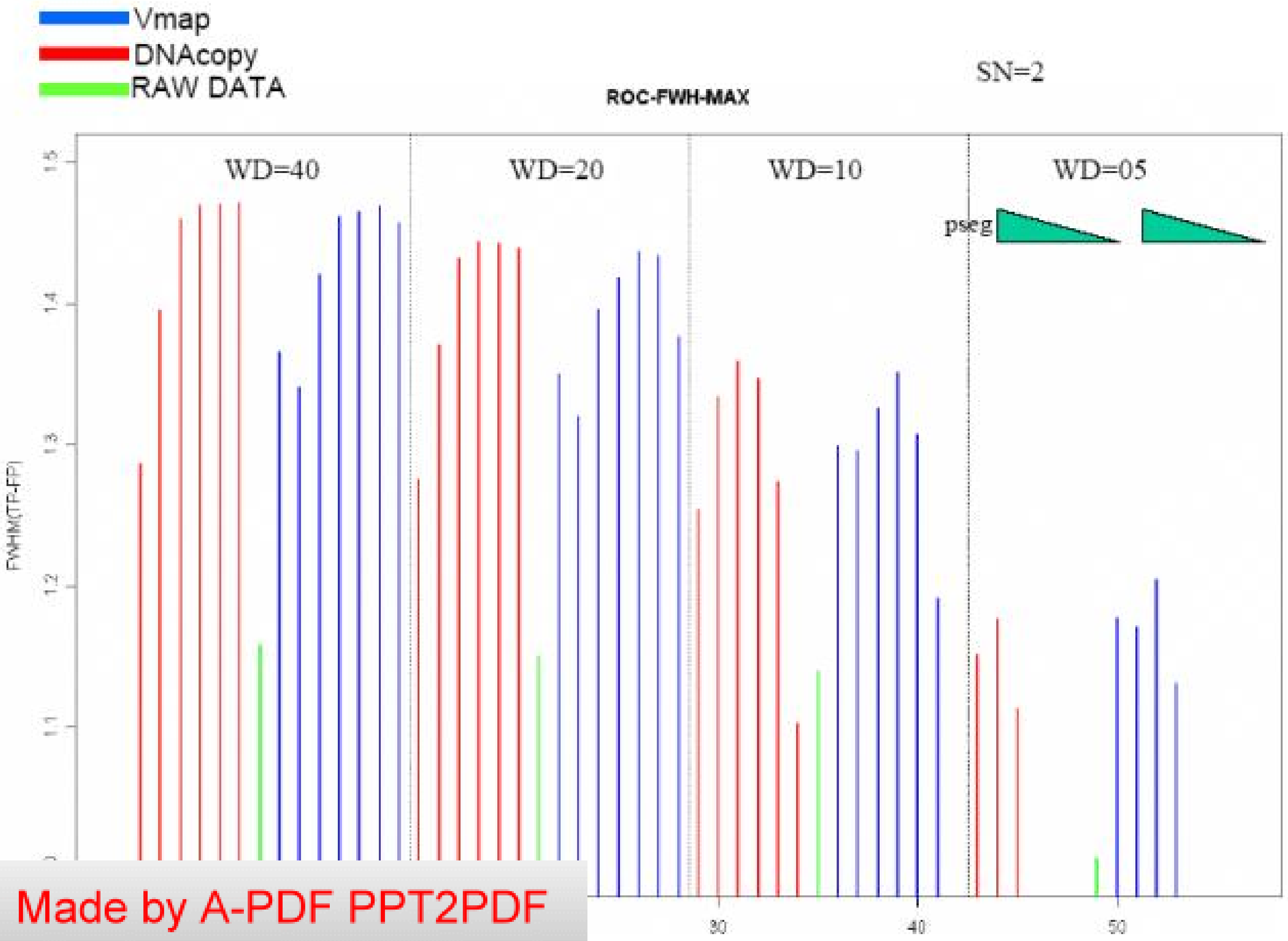


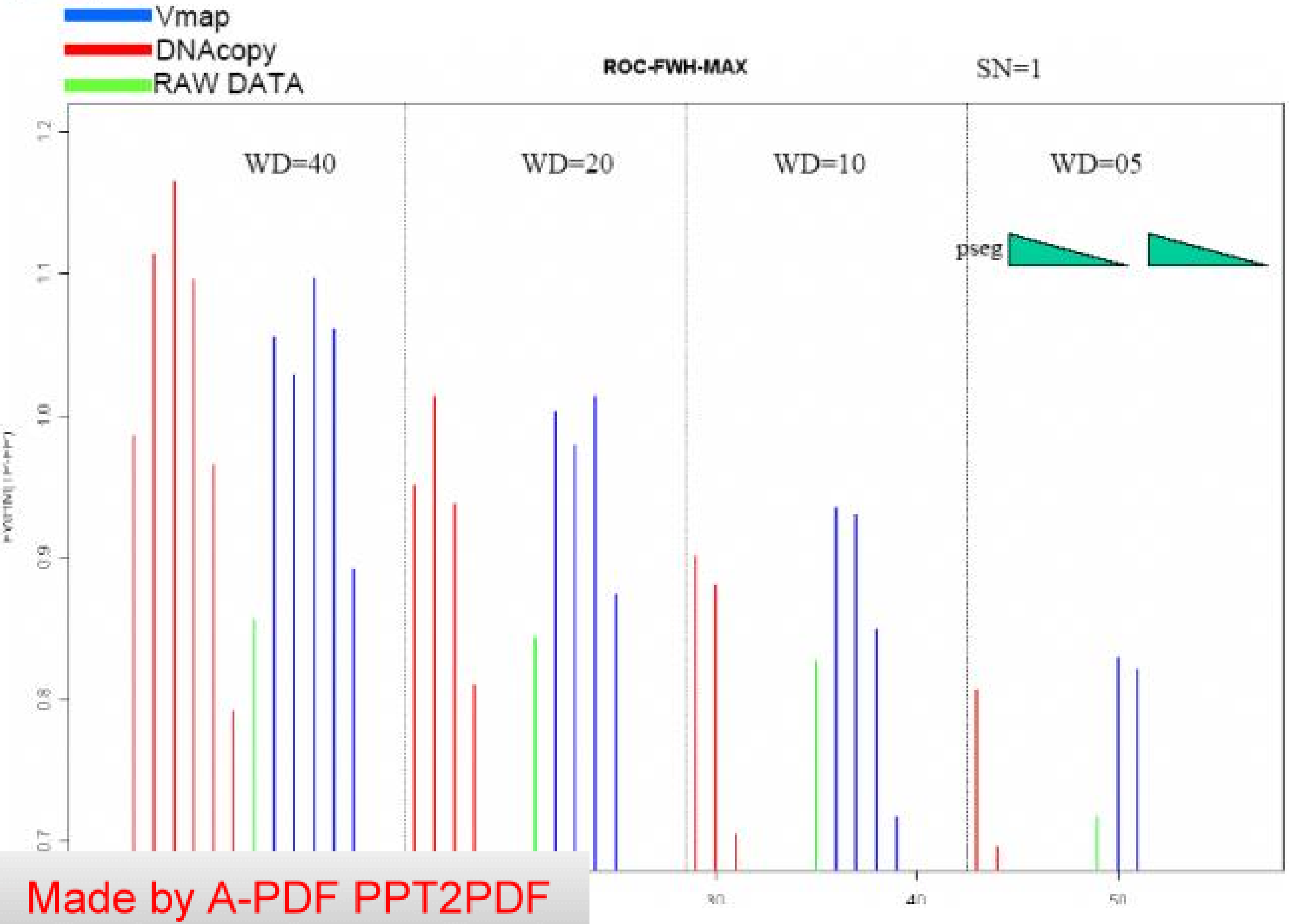
SN=4

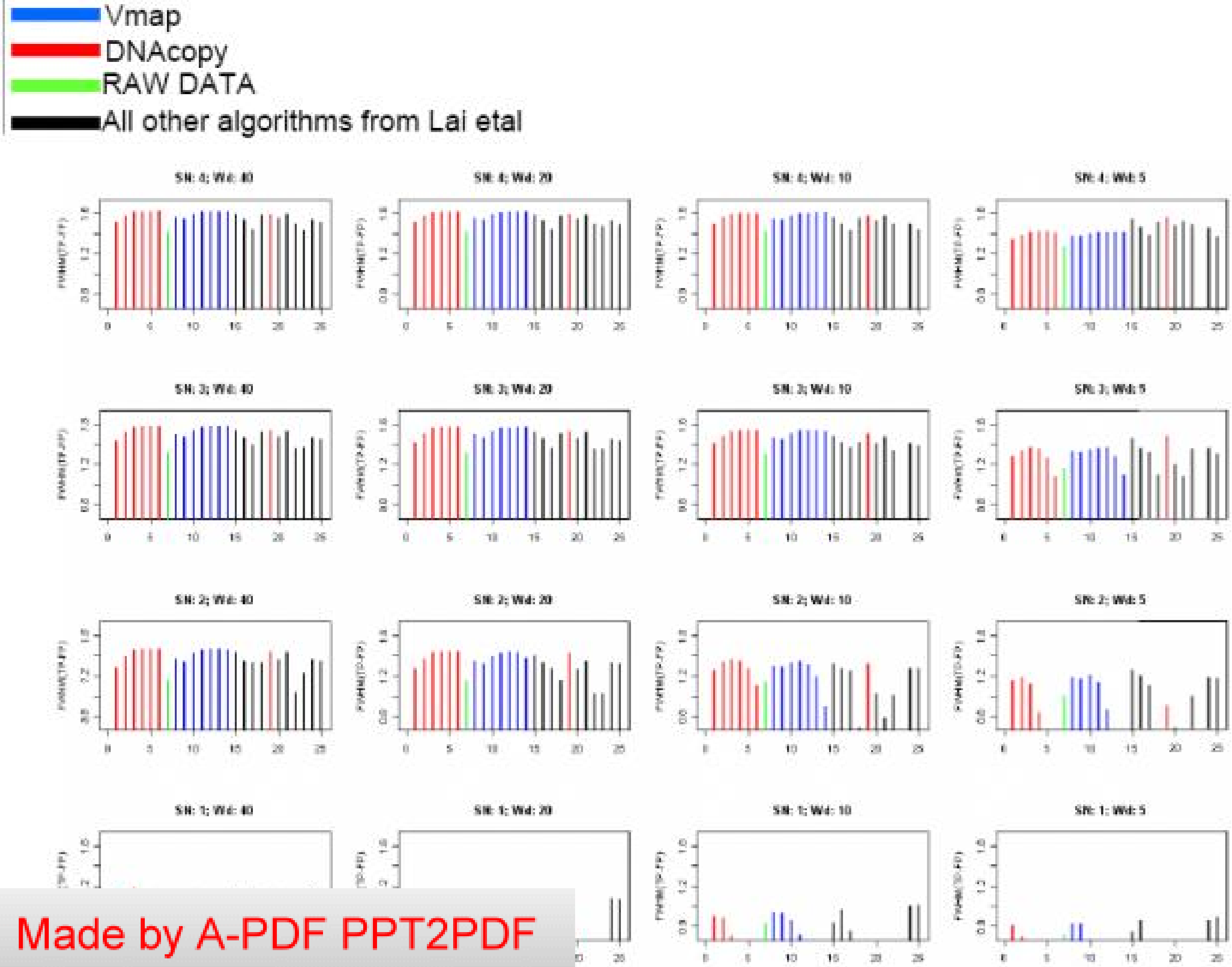

Vmap
DNAcopy
RAW DATA
ROC-FWH-MAX
WD=40
WD=20
WD=10
WD=05
pseg
FWHM(TP-FP)
1.70
1.65
1.60
1.55
1.50
1.45
1.40
20
30
40



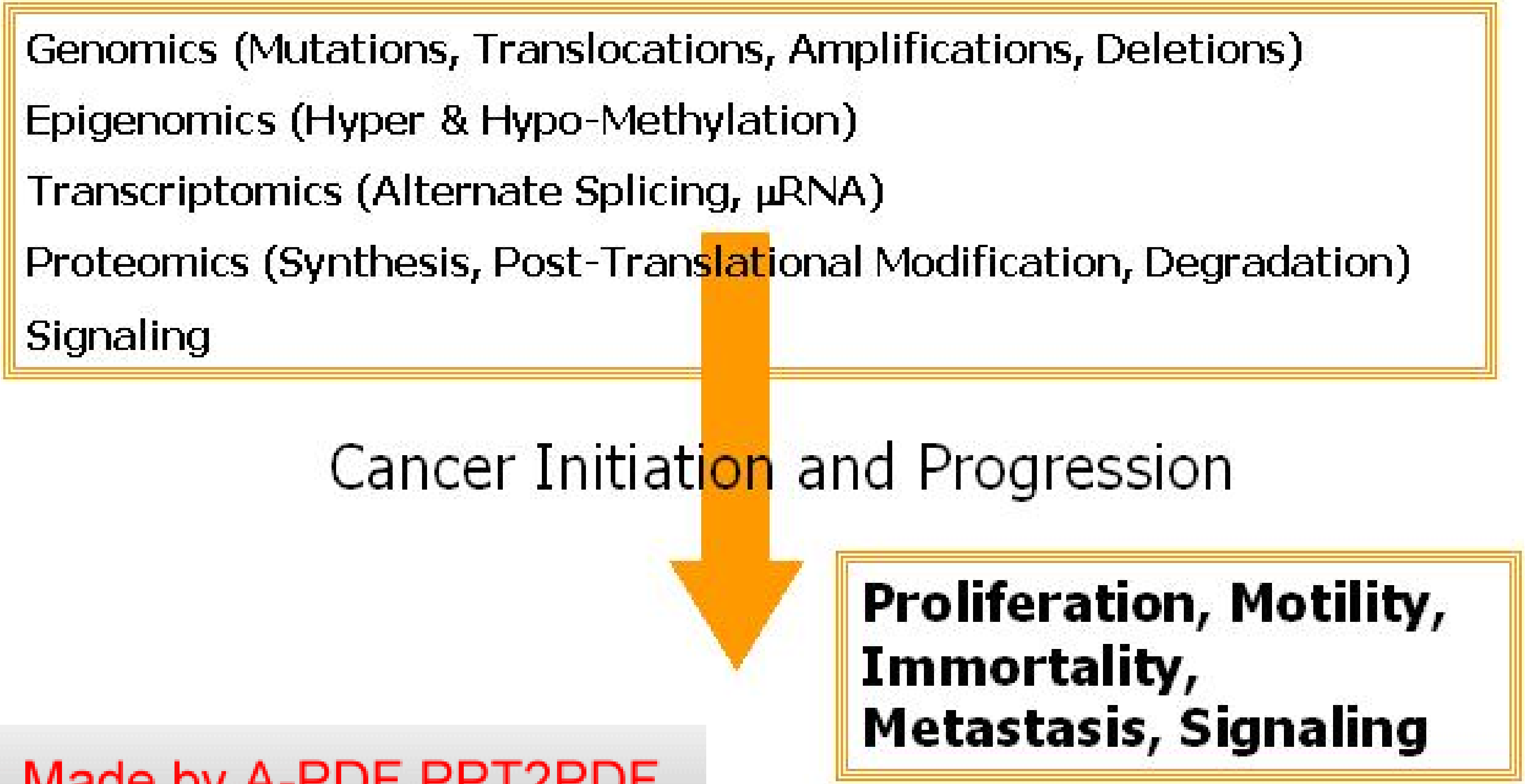
Vmap
DNAcopy
RAW DATA
SN=3
ROC-FWH-MAX
WD=40
WD=20
WD=10
WD=05
pseg
FWHM(TP-FP)
1.6
1.5
1.4
1.3
1.2
1.1
1.0
30
40
50


Vmap
DNAcopy
RAW DATA
SN=2
ROC-FWH-MAX
WD=40
WD=20
WD=10
WD=05
pseg
FWHM(TP-FP)
1.5
1.4
1.3
1.2
1.1
30
40
50


Vmap
DNAcopy
RAW DATA
ROC-FWH-MAX
SN=1
WD=40
WD=20
WD=10
WD=05
pseg
1.2
1.1
1.0
0.9
0.8
0.7
30
40
50


Vmap
DNAcopy
RAW DATA
All other algorithms from Lai etal
SN: 4; Wd: 40
SN: 4; Wd: 20
SN: 4; Wd: 10
SN: 4; Wd: 5
SN: 3; Wd: 40
SN: 3; Wd: 20
SN: 3; Wd: 10
SN: 3; Wd: 5
SN: 2; Wd: 40
SN: 2; Wd: 20
SN: 2; Wd: 10
SN: 2; Wd: 5
SN: 1; Wd: 40
SN: 1; Wd: 20
SN: 1; Wd: 10
SN: 1; Wd: 5
FWHM(TP-FP)
0.8
1.2
1.6
0
5
10
15
20
25

Cancer Initiation and Progression



Genomics (Mutations, Translocations, Amplifications, Deletions)

Epigenomics (Hyper & Hypo-Methylation)

Transcriptomics (Alternate Splicing, µRNA)

Proteomics (Synthesis, Post-Translational Modification, Degradation)

Signaling

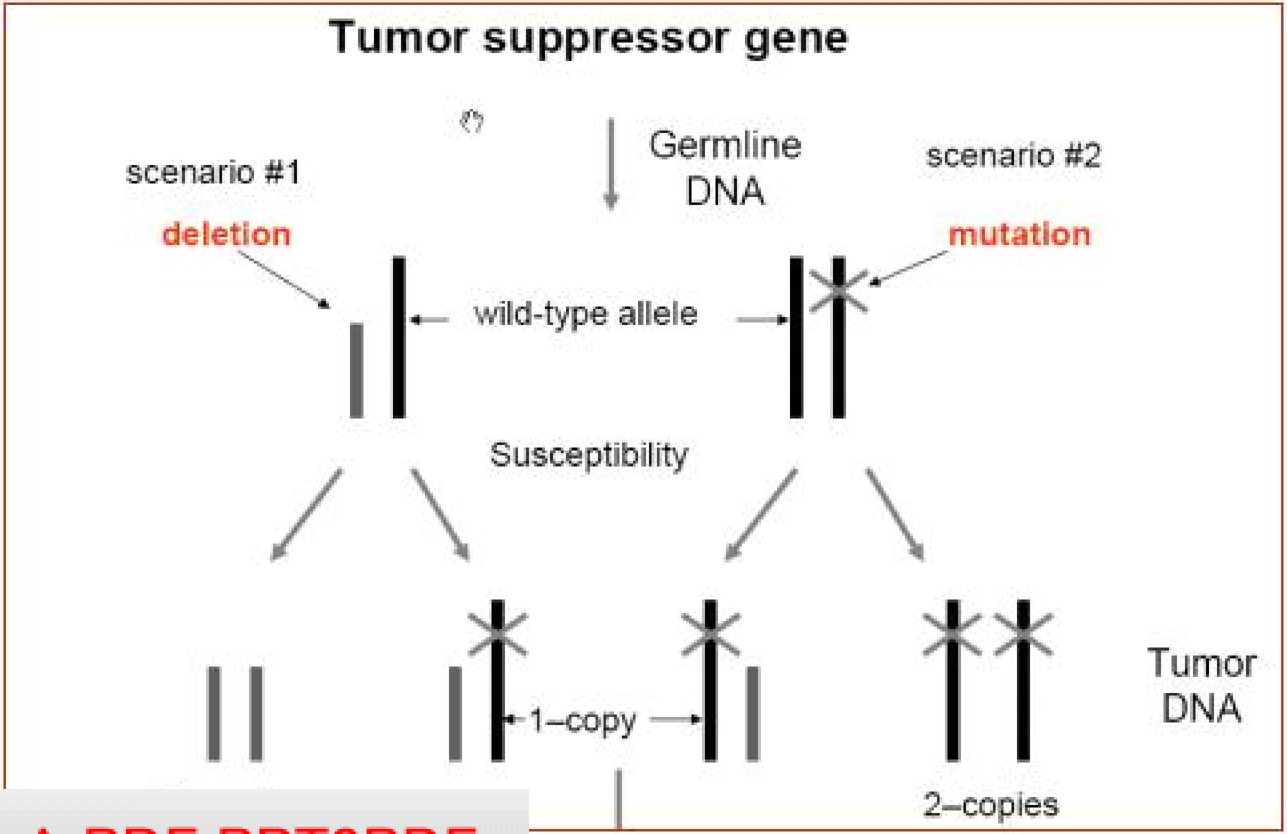
Cancer Initiation and Progression

Proliferation, Motility, Immortality, Metastasis, Signaling

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.

Genetics of LOH



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.

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

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

where A is the event “I amplified’’ (for oncogenes) and [illegible] suppressor genes).

Relative Risk Score (cont)

- The first part can be estimated from data:

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

- 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

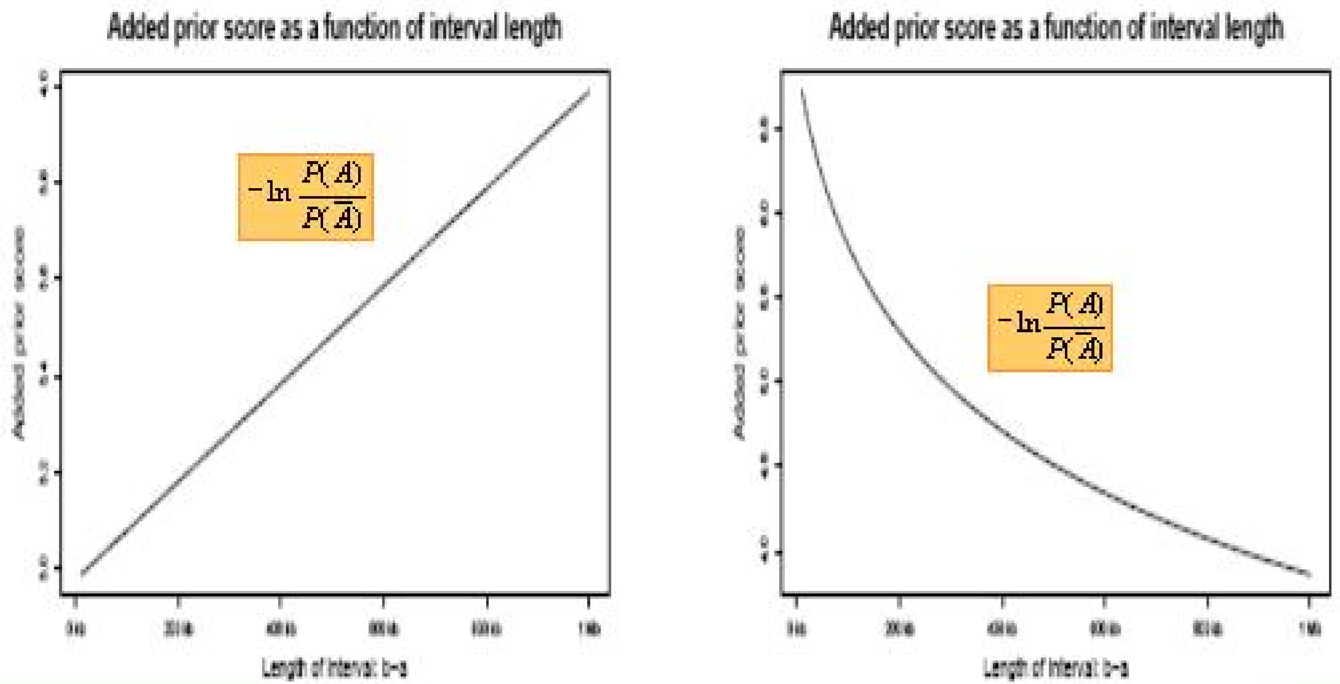
Relative Risk Score: Marginal

- 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.

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

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

Prior Score



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:

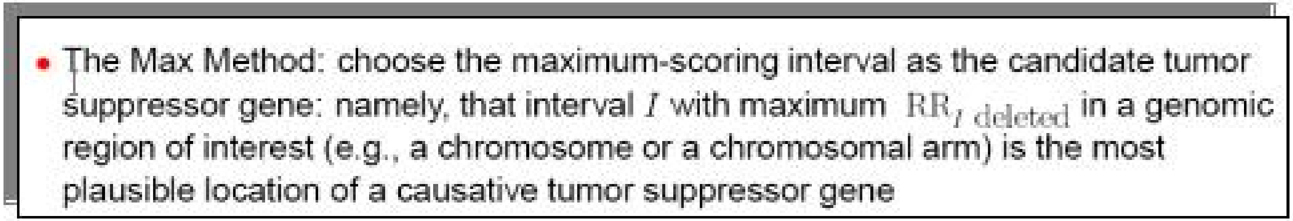
  The Max method &

  The LR (left-right) method

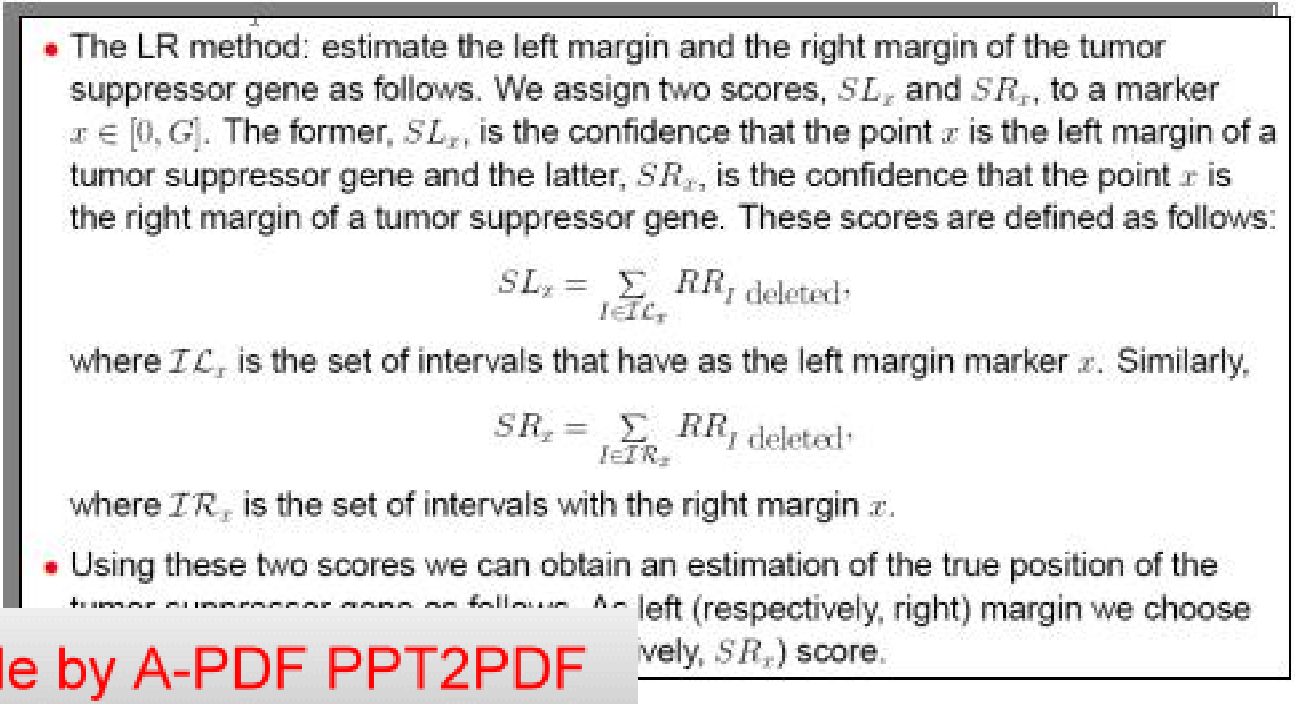
High Scoring Intervals

- 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.

Finding the TSG



- The Max Method: choose the maximum-scoring interval as the candidate tumor suppressor gene: namely, that interval I with maximum $\mathrm{RR}_{I\ \mathrm{deleted}}$ in a genomic region of interest (e.g., a chromosome or a chromosomal arm) is the most plausible location of a causative tumor suppressor gene



- The LR method: estimate the left margin and the right margin of the tumor suppressor gene as follows. We assign two scores, SL_x and SR_x, to a marker $x \in [0, G]$. The former, SL_x, is the confidence that the point x is the left margin of a tumor suppressor gene and the latter, SR_x, is the confidence that the point x is the right margin of a tumor suppressor gene. These scores are defined as follows:

$$SL_x = \sum_{I \in \mathcal{IL}_x} RR_{I\ \mathrm{deleted}},$$

where $\mathcal{IL}_x$ is the set of intervals that have as the left margin marker x. Similarly,

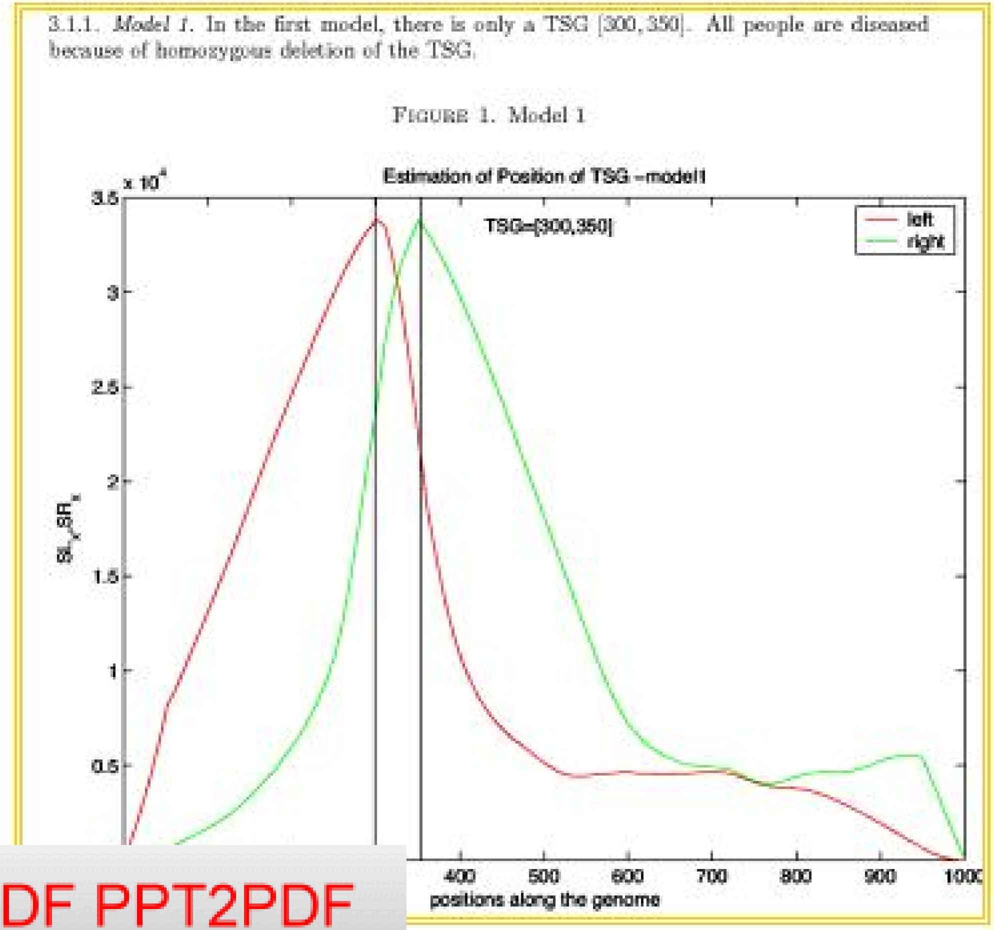
$$SR_x = \sum_{I \in \mathcal{IR}_x} RR_{I\ \mathrm{deleted}},$$

where $\mathcal{IR}_x$ is the set of intervals with the right margin x.

- Using these two scores we can obtain an estimation of the true position of the tumor suppressor gene as follows. As left (respectively, right) margin we choose

vely, SR_x) score.

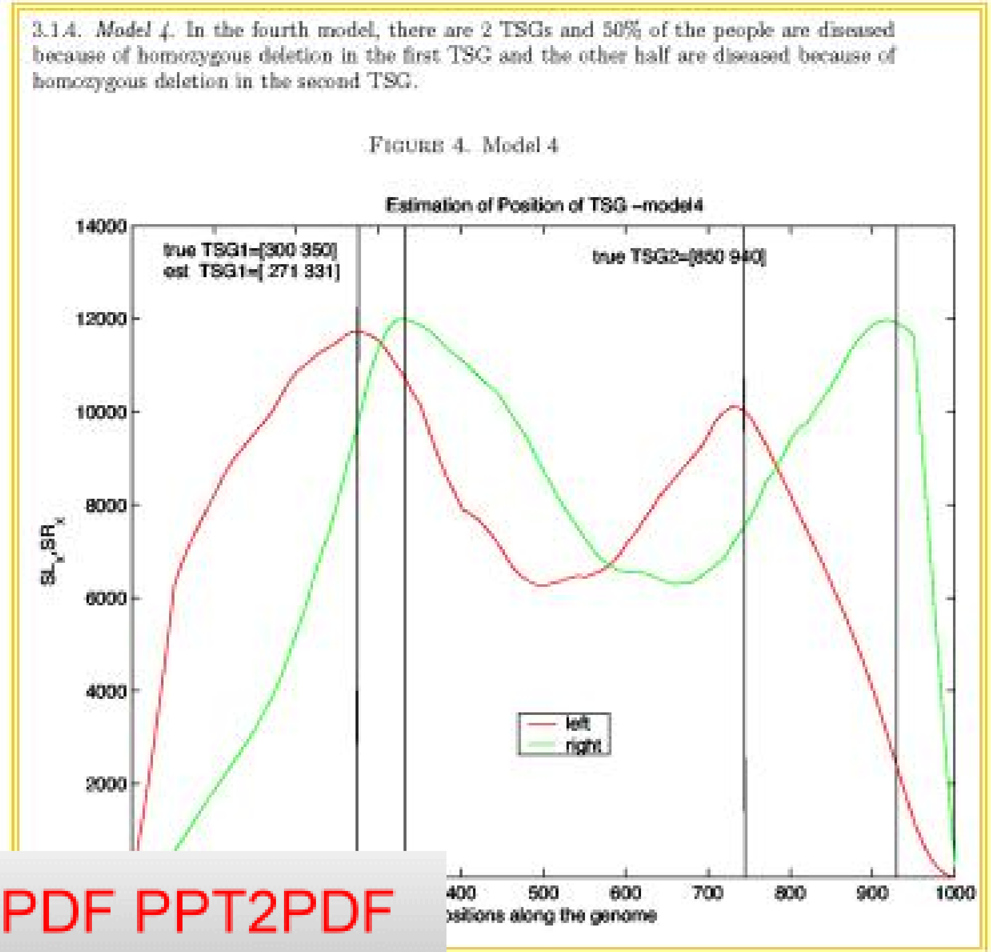
Model Simulation



3.1.1. *Model 1.* In the first model, there is only a TSG [300, 350]. All people are diseased because of homozygous deletion of the TSG.

FIGURE 1. Model 1

Model Simulation



3.1.4. *Model 4.* In the fourth model, there are 2 TSGs and 50% of the people are diseased because of homozygous deletion in the first TSG and the other half are diseased because of homozygous deletion in the second TSG.

FIGURE 4. Model 4

Significance Testing

- We now know how to estimate the most likely location of a cancer gene in a genomic region of interest.
  - **Call the interval 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 l_{tsg} is a tumor suppressor gene, then its neighborhood should contain an unusually large number of breakpoints, signifying a region with many deletions

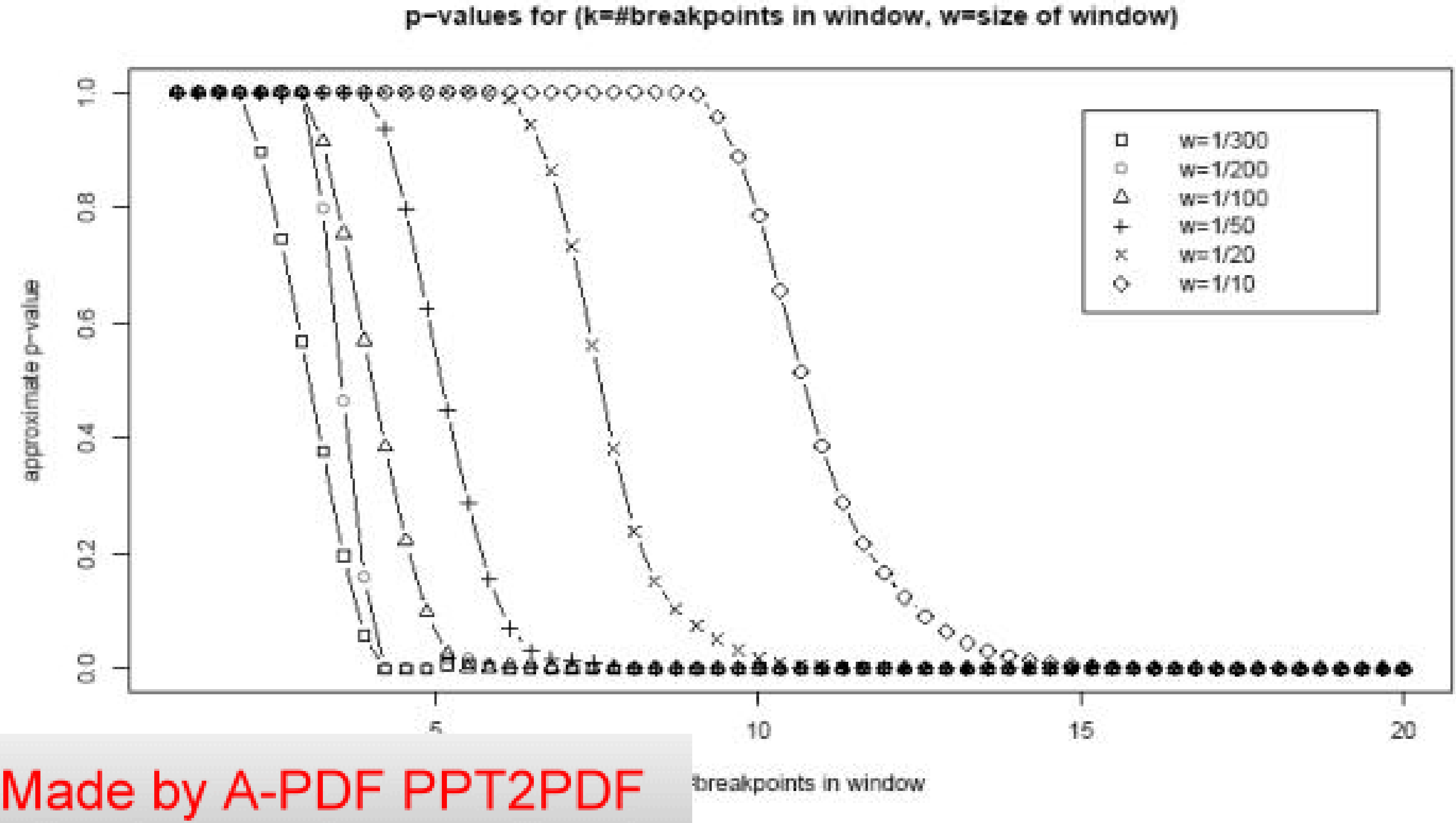
Scan Statistics

- 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

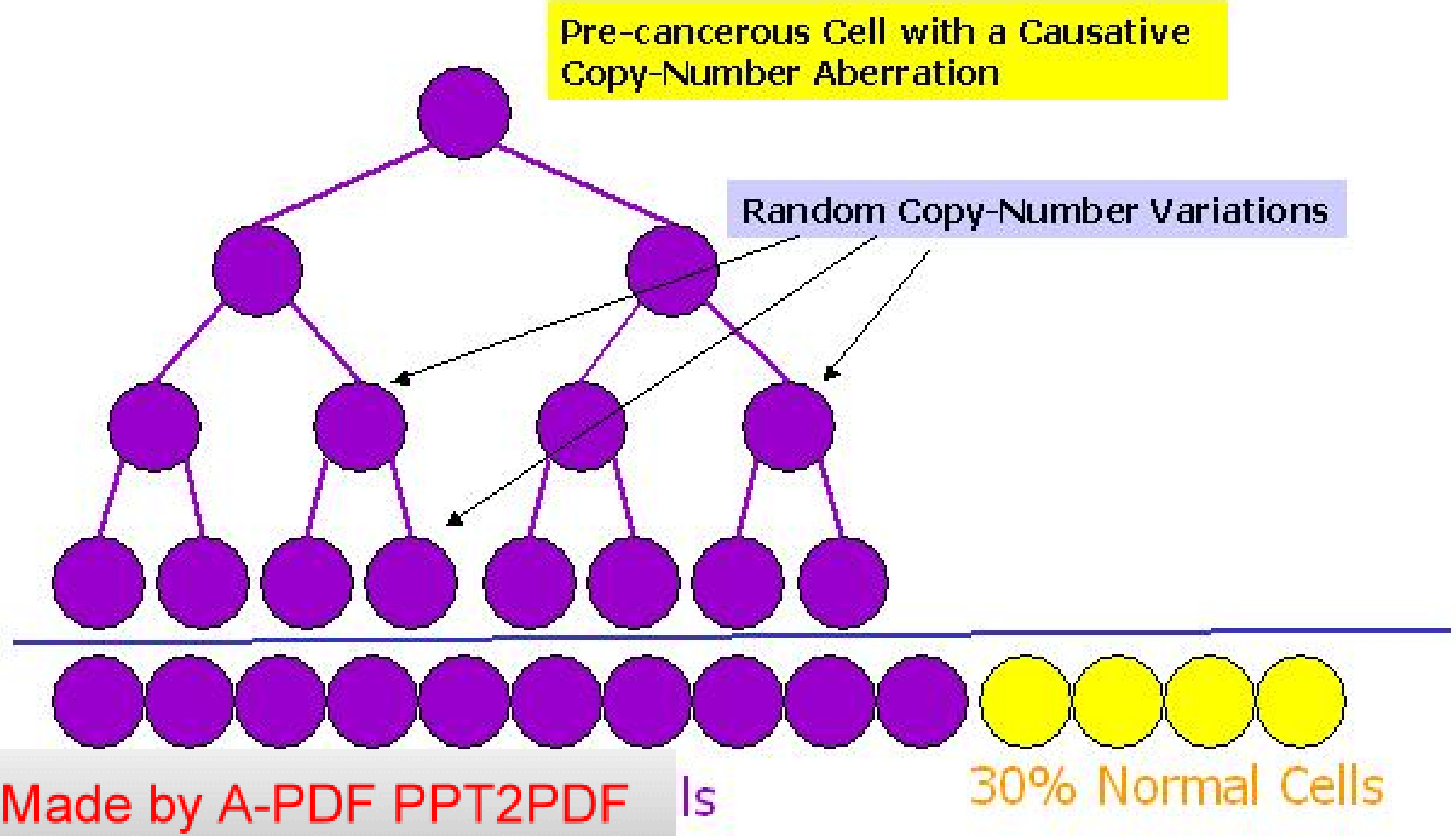
$$P(S_w \geq k) \approx (kw^{-1} - N - 1)b(k;N,w) + 2G_b(k;N,w) \quad \textbf{where}$$

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

Scan Statistics



Simulation Model



Pre-cancerous Cell with a Causative Copy-Number Aberration

Random Copy-Number Variations

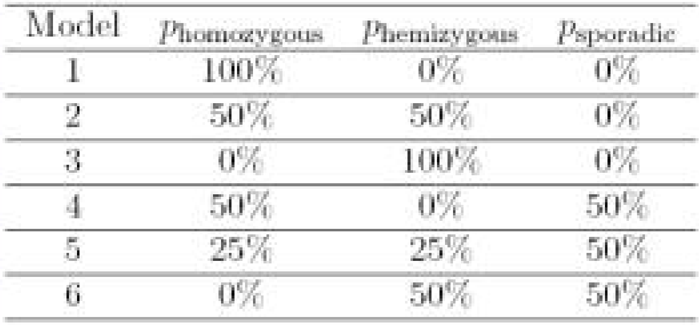
30% Normal Cells

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:

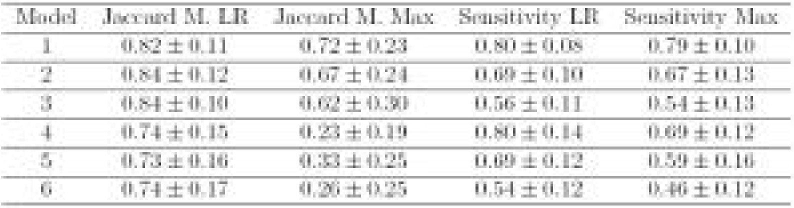
$$J(E,T) = \frac{|E \cap T|}{|E \cup T|}$$

Results



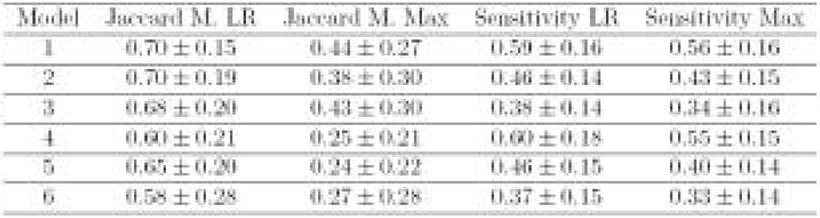
Model	$P_{homozygous}$	$P_{hemizygous}$	$P_{sporadic}$
1	100%	0%	0%
2	50%	50%	0%
3	0%	100%	0%
4	50%	0%	50%
5	25%	25%	50%
6	0%	50%	50%

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.

Results

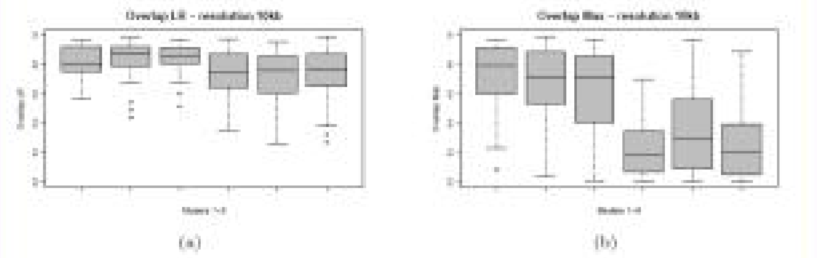


Figure 2: Boxplots of the Jaccard measure of overlap for each of the six models. Average inter-marker distance is 10 Kb. (a) LR, (b) Max.

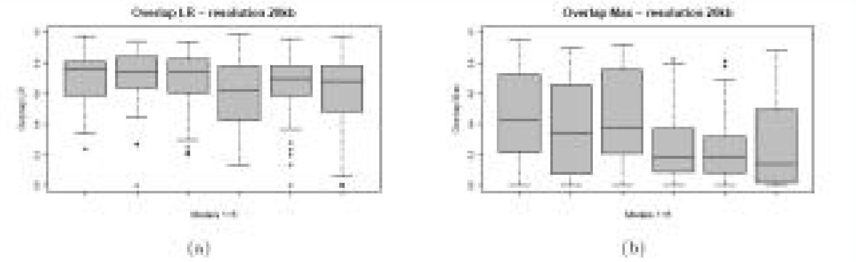


Figure 4: Boxplots of the Jaccard measure of overlap for each of the six models. Average inter-marker distance is 20 Kb. (a) LR, (b) Max.

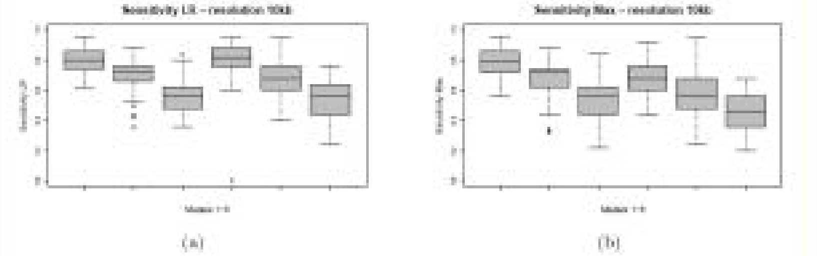


Figure 3: Boxplots of the sensitivity measure for each of the six models. Average inter-marker distance is 10 Kb. (a) LR, (b) Max.

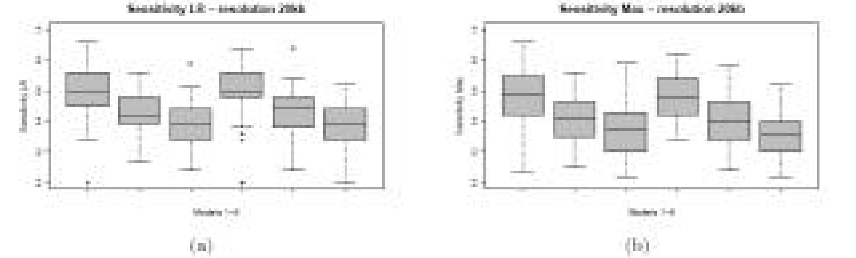


Figure 5: Boxplots of the sensitivity measure for each of the six models. Average inter-marker distance is 20 Kb. (a) LR, (b) Max.

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).

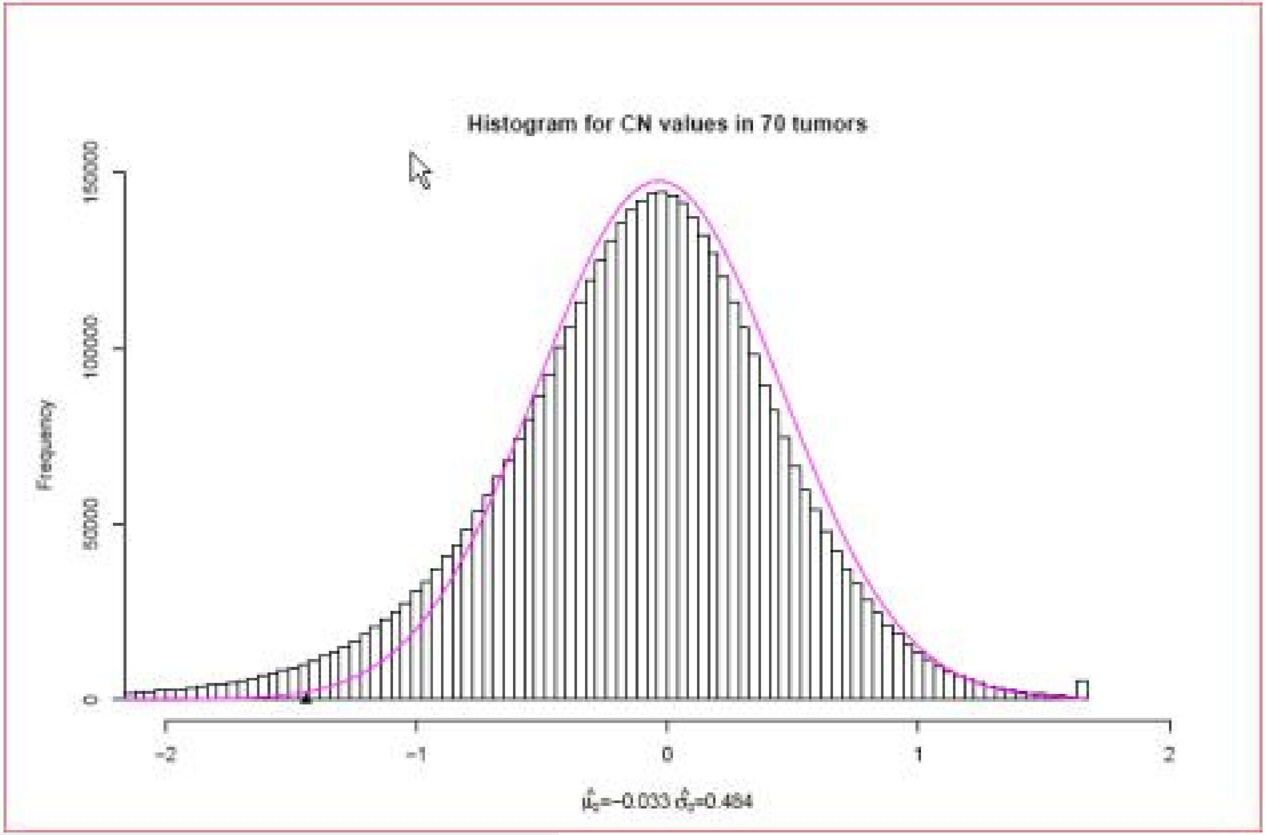
TSG

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*
[illegible]	Del Lung Cancer

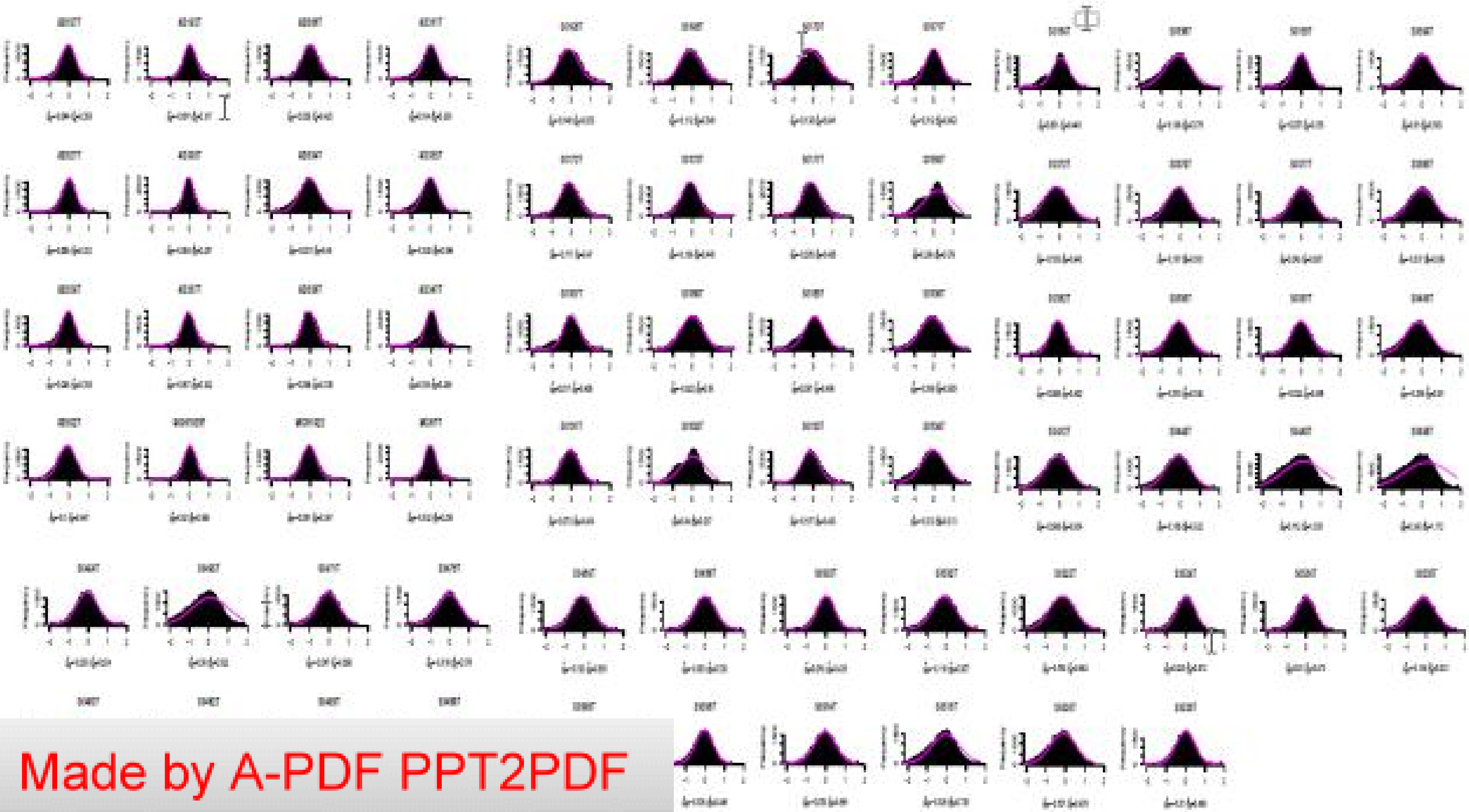
Onco-gene

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)

Copy Number Data



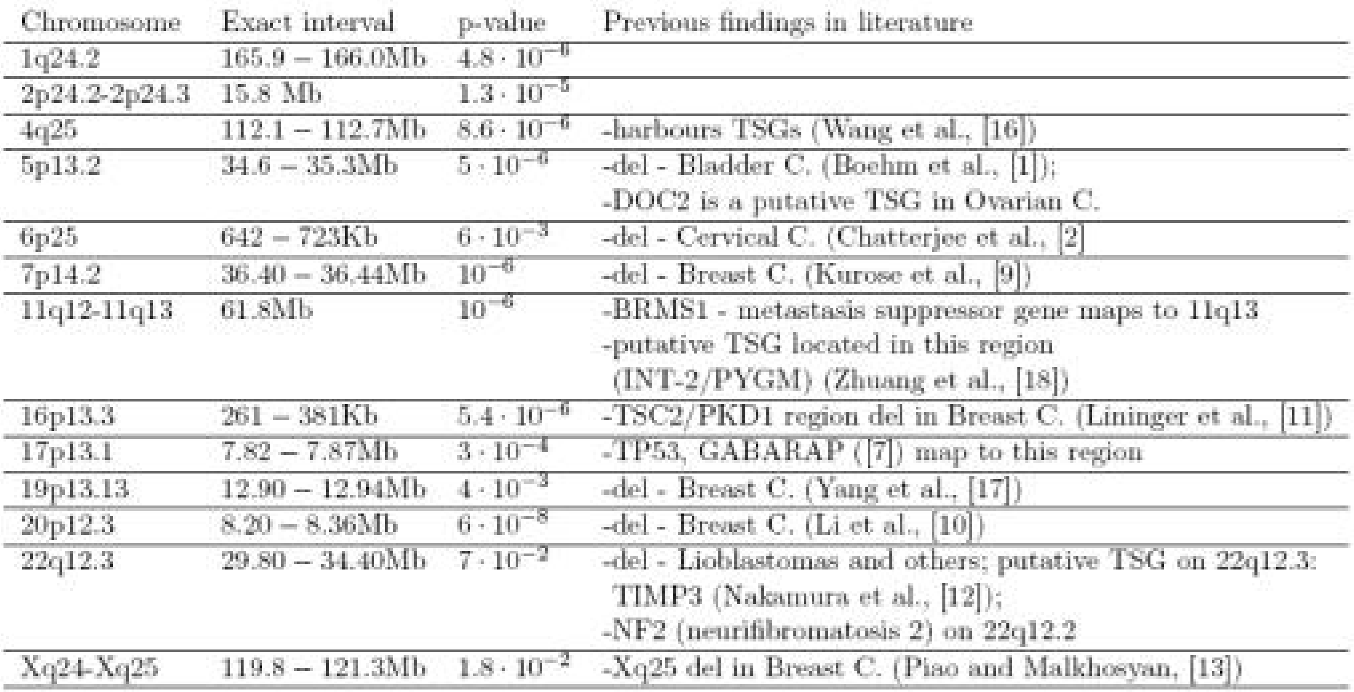
All 70 Patients



Breast Cancer Dataset

- We tested our algorithm on a real dataset on breast cancer (Pollack et. al 2002).
  - The dataset consists of DNA copy number changes across 6,691 human genes in 44 primary breast tumors and 15 breast cancer cell lines.
  - We detected several interesting regions, many of which have previously been implicated in breast cancer or other cancers.

Results- Breast Cancer



Chromosome	Exact interval	p-value	Previous findings in literature
1q24.2	165.9 – 166.0Mb	$4.8 \cdot 10^{-6}$	
2p24.2-2p24.3	15.8 Mb	$1.3 \cdot 10^{-5}$	
4q25	112.1 – 112.7Mb	$8.6 \cdot 10^{-6}$	-harbours TSGs (Wang et al., [16])
5p13.2	34.6 – 35.3Mb	$5 \cdot 10^{-6}$	-del - Bladder C. (Boehm et al., [1]); -DOC2 is a putative TSG in Ovarian C.
6p25	642 – 723Kb	$6 \cdot 10^{-3}$	-del - Cervical C. (Chatterjee et al., [2]
7p14.2	36.40 – 36.44Mb	10^{-6}	-del - Breast C. (Kurose et al., [9])
11q12-11q13	61.8Mb	10^{-6}	-BRMS1 - metastasis suppressor gene maps to 11q13 -putative TSG located in this region (INT-2/PYGM) (Zhuang et al., [18])
16p13.3	261 – 381Kb	$5.4 \cdot 10^{-6}$	-TSC2/PKD1 region del in Breast C. (Lininger et al., [11])
17p13.1	7.82 – 7.87Mb	$3 \cdot 10^{-4}$	-TP53, GABARAP ([7]) map to this region
19p13.13	12.90 – 12.94Mb	$4 \cdot 10^{-3}$	-del - Breast C. (Yang et al., [17])
20p12.3	8.20 – 8.36Mb	$6 \cdot 10^{-8}$	-del - Breast C. (Li et al., [10])
22q12.3	29.80 – 34.40Mb	$7 \cdot 10^{-2}$	-del - Lioblastomas and others; putative TSG on 22q12.3: TIMP3 (Nakamura et al., [12]); -NF2 (neurifibromatosis 2) on 22q12.2
Xq24-Xq25	119.8 – 121.3Mb	$1.8 \cdot 10^{-2}$	-Xq25 del in Breast C. (Piao and Malkhosyan, [13])

Table 4: Significant Deleted Regions on the Real Dataset. Pointwise p-values are given for each region. In order to obtain a chromosomal p-value of 0.05 using a Bonferroni ultiplied by 10.

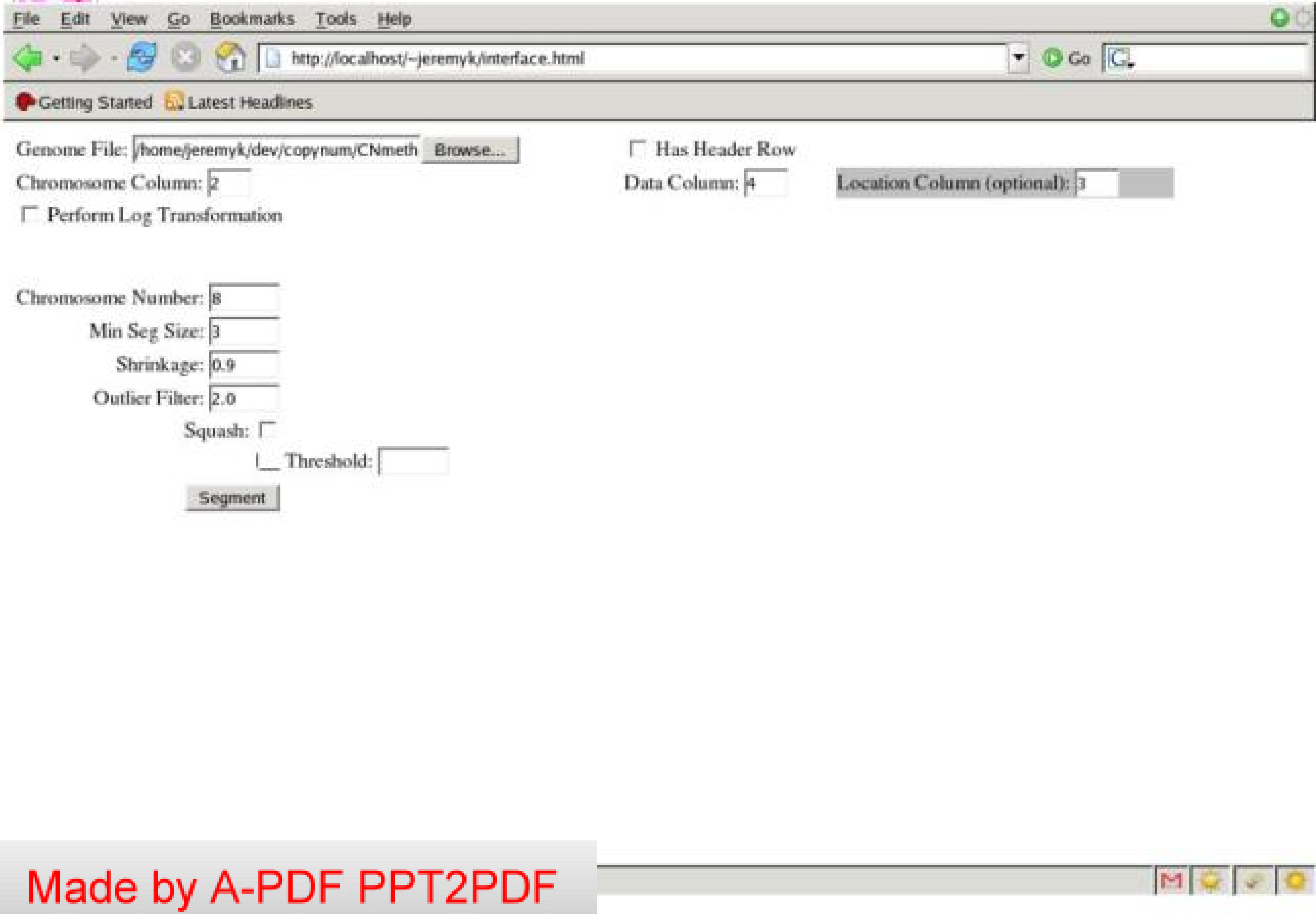
Extensions

- Combining with Gene Expression data
  - Uses a pathway model (Inferred with a Truncated Stein Shrinkage)
  - Uses Scan-statistics on a graph to determine genes associated with CNVs, cascades in pathways, & “Others” (mutations, translocations or epigenomics)
- Generalizes to other datasets (e.g. proteomics etc.)

Software Architecture

Current Implementation

- NYU Array CGH
- Agnostic to Technology:
  - Works well for BAC array, ROMA, Agilent, Affymetrix
  - Raw Affymetrix data is noisier, but our new algorithm for "background correction and summarization (BCS)" makes Affymetrix-data significantly better.
- Scalable: Fast implementation, with visualization and integration (**Publicly Available**)
- Generalized Global Analysis (LOH analysis, detecting TSG and oncogenes)


File Edit View Go Bookmarks Tools Help
http://localhost/~jeremyk/interface.html
Go
Getting Started
Latest Headlines
Browse...

Genome File: /home/jeremyk/dev/copynum/CNmeth

□ Has Header Row

Chromosome Column: 2

Data Column: 4

Location Column (optional): 3

□ Perform Log Transformation

Chromosome Number: 8

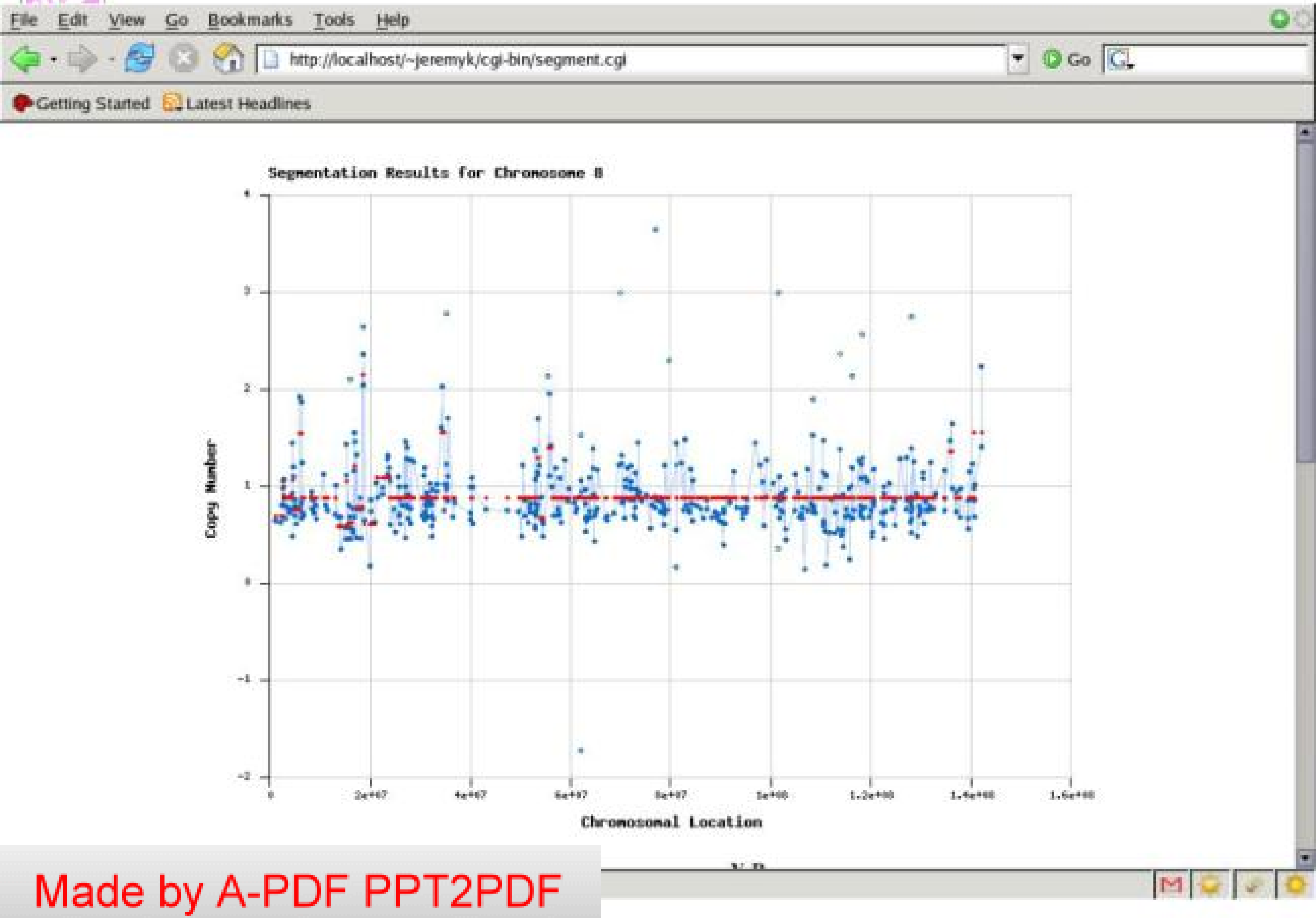
Min Seg Size: 3

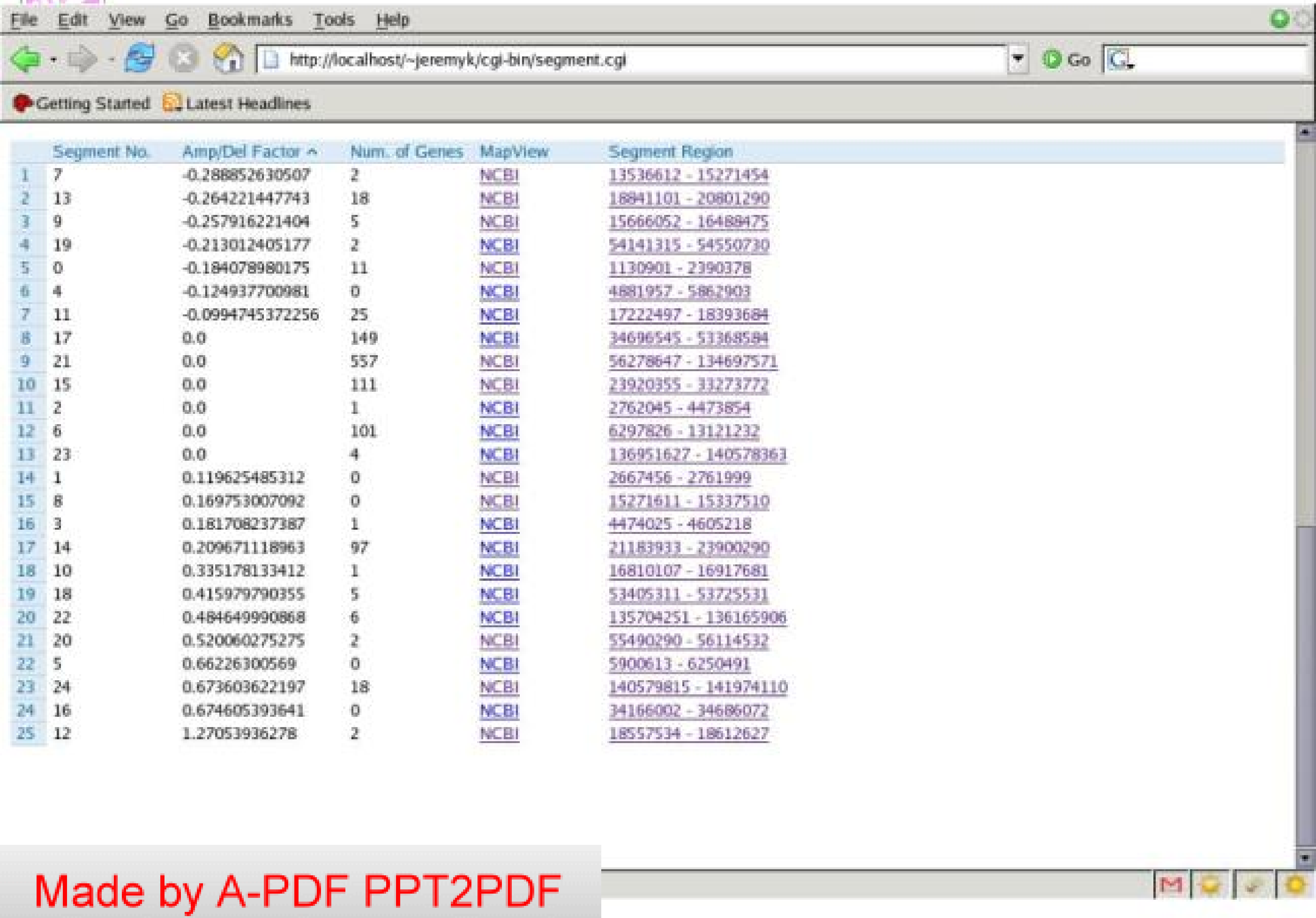
Shrinkage: 0.9

Outlier Filter: 2.0

Squash: □

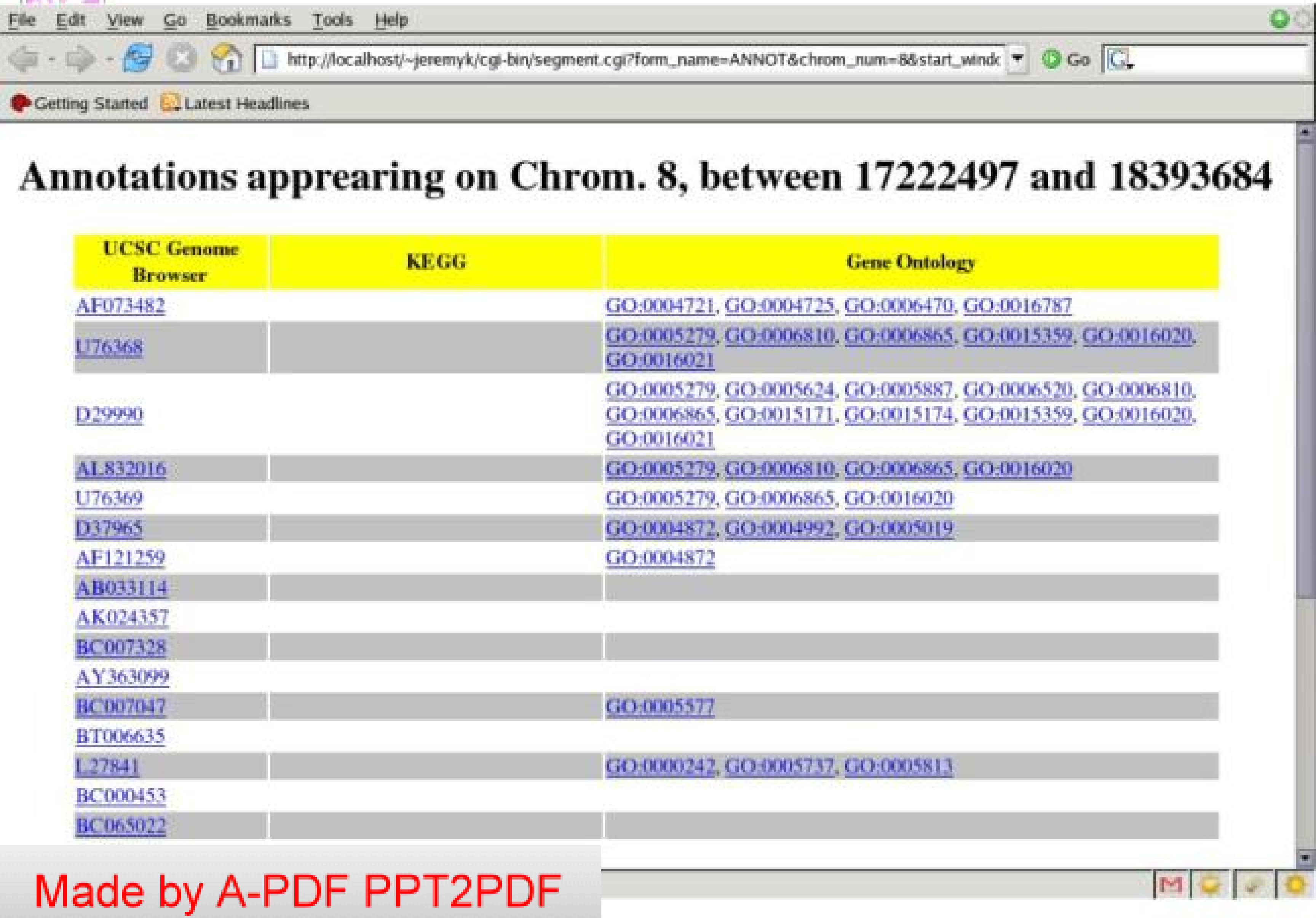
Threshold:


File Edit View Go Bookmarks Tools Help
http://localhost/~jeremyk/cgi-bin/segment.cgi
Go
Getting Started
Latest Headlines
Segmentation Results for Chromosome 8
Copy Number
4
3
2
1
0
-1
-2
0
2e+07
4e+07
6e+07
8e+07
1e+08
1.2e+08
1.4e+08
1.6e+08
Chromosomal Location



http://localhost/~jeremyk/cgi-bin/segment.cgi

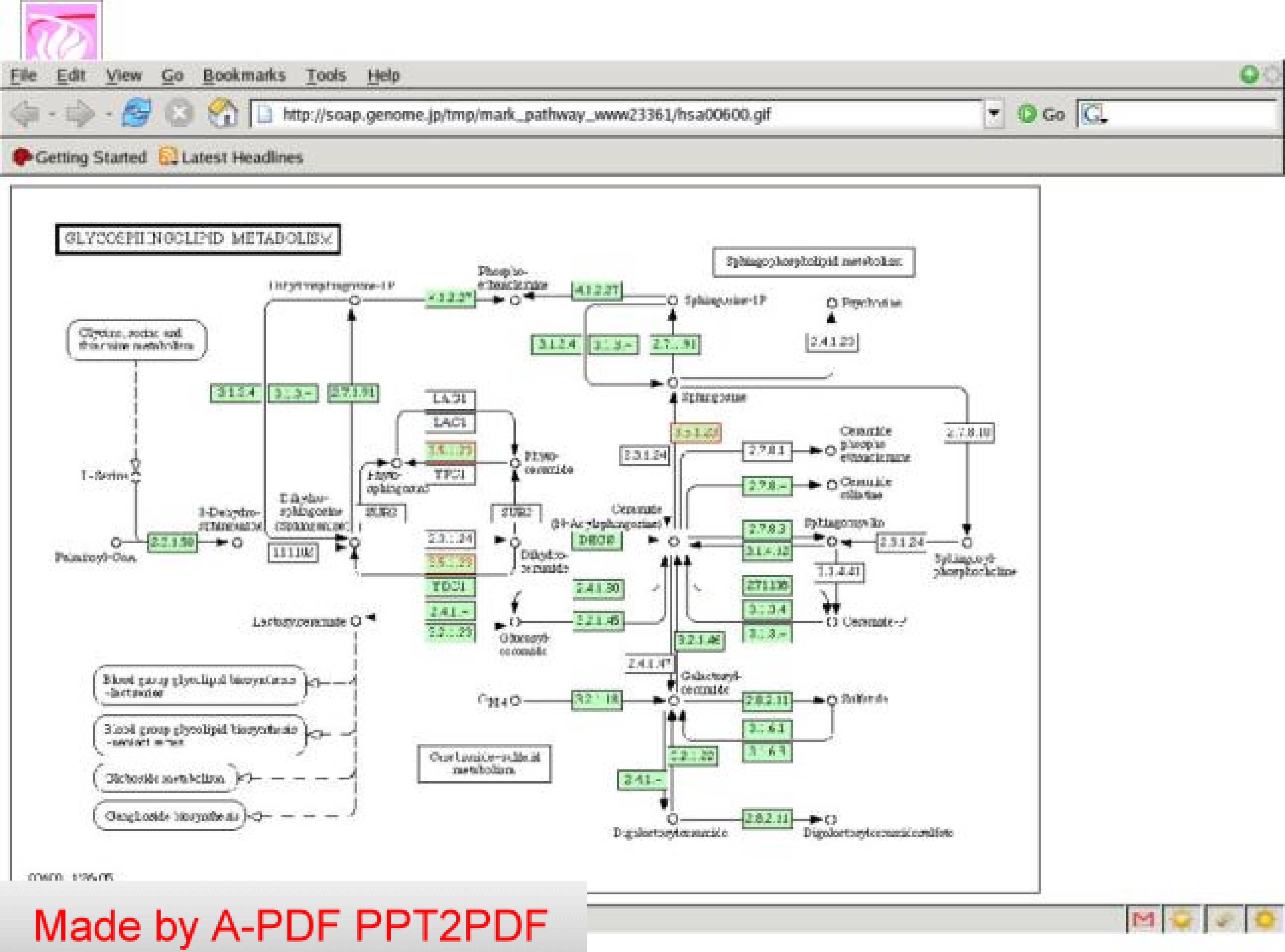
	Segment No.	Amp/Del Factor ^	Num. of Genes	MapView	Segment Region
1	7	-0.288852630507	2	NCBI	13536612 - 15271454
2	13	-0.264221447743	18	NCBI	18841101 - 20801290
3	9	-0.257916221404	5	NCBI	15666052 - 16488475
4	19	-0.213012405177	2	NCBI	54141315 - 54550730
5	0	-0.184078980175	11	NCBI	1130901 - 2390378
6	4	-0.124937700981	0	NCBI	4881957 - 5862903
7	11	-0.0994745372256	25	NCBI	17222497 - 18393684
8	17	0.0	149	NCBI	34696545 - 53368584
9	21	0.0	557	NCBI	56278647 - 134697571
10	15	0.0	111	NCBI	23920355 - 33273772
11	2	0.0	1	NCBI	2762045 - 4473854
12	6	0.0	101	NCBI	6297826 - 13121232
13	23	0.0	4	NCBI	136951627 - 140578363
14	1	0.119625485312	0	NCBI	2667456 - 2761999
15	8	0.169753007092	0	NCBI	15271611 - 15337510
16	3	0.181708237387	1	NCBI	4474025 - 4605218
17	14	0.209671118963	97	NCBI	21183933 - 23900290
18	10	0.335178133412	1	NCBI	16810107 - 16917681
19	18	0.415979790355	5	NCBI	53405311 - 53725531
20	22	0.484649990868	6	NCBI	135704251 - 136165906
21	20	0.520060275275	2	NCBI	55490290 - 56114532
22	5	0.66226300569	0	NCBI	5900613 - 6250491
23	24	0.673603622197	18	NCBI	140579815 - 141974110
24	16	0.674605393641	0	NCBI	34166002 - 34686072
25	12	1.27053936278	2	NCBI	18557534 - 18612627



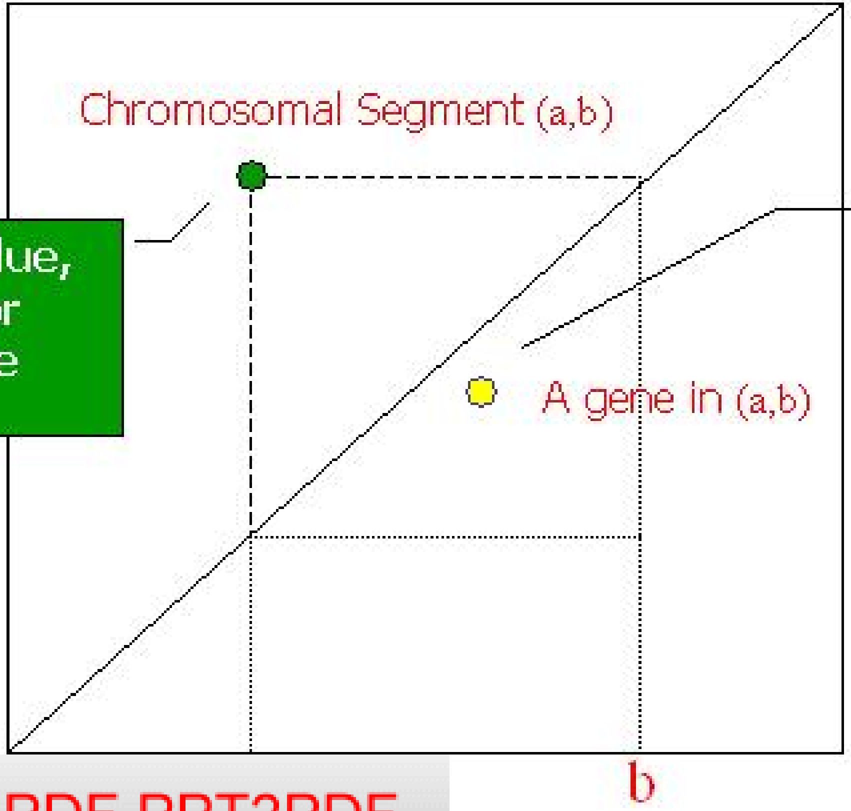
http://localhost/~jeremyk/cgi-bin/segment.cgi?form_name=ANNOT&chrom_num=8&start_windc

Annotations apprearing on Chrom. 8, between 17222497 and 18393684

UCSC Genome Browser	KEGG	Gene Ontology
AF073482		GO:0004721, GO:0004725, GO:0006470, GO:0016787
U76368		GO:0005279, GO:0006810, GO:0006865, GO:0015359, GO:0016020, GO:0016021
D29990		GO:0005279, GO:0005624, GO:0005887, GO:0006520, GO:0006810, GO:0006865, GO:0015171, GO:0015174, GO:0015359, GO:0016020, GO:0016021
AL832016		GO:0005279, GO:0006810, GO:0006865, GO:0016020
U76369		GO:0005279, GO:0006865, GO:0016020
D37965		GO:0004872, GO:0004992, GO:0005019
AF121259		GO:0004872
AB033114		
AK024357		
BC007328		
AY363099		
BC007047		GO:0005577
BT006635		
L27841		GO:0000242, GO:0005737, GO:0005813
BC000453		
BC065022		


File
Edit
View
Go
Bookmarks
Tools
Help
http://soap.genome.jp/tmp/mark_pathway_www23361/hsa00600.gif
Go
Getting Started
Latest Headlines
GLYCOSPHINGOLIPID METABOLISM
Sphingophospholipid metabolism
Glycine, serine and threonine metabolism
Dihydrosphingosine-1P
Phospho-ethanolamine
4.1.2.27
4.1.2.27
Sphingosine-1P
Psychosine
3.1.3.4
3.1.3.-
2.7.1.91
2.4.1.23
3.1.3.4
3.1.3.-
2.7.1.91
Sphingosine
LAG1
LAC1
3.5.1.23
2.7.8.10
3.5.1.23
TSC1
Phyto-sphingosine
Phyto-ceramide
2.3.1.24
2.7.8.1
Ceramide phospho-ethanolamine
2.7.8.-
Ceramide choline
L-Serine
3-Dehydro-sphinganine
Dihydro-sphingosine (Sphinganine)
SUR2
SUR2
Ceramide (N-Acylsphingosine)
Sphingomyelin
2.7.8.3
3.1.4.12
2.3.1.24
Sphingosyl-phosphocholine
Palmitoyl-CoA
2.3.1.50
1.1.1.102
2.3.1.24
3.5.1.23
DEGS
Dihydro-ceramide
YDC1
2.4.1.80
3.1.4.41
2.7.1.138
3.1.3.4
3.1.3.-
Ceramide-1P
2.4.1.-
3.2.1.45
Lactosylceramide
3.2.1.23
Glucosyl-ceramide
3.2.1.46
2.4.1.47
Galactosyl-ceramide
Blood group glycolipid biosynthesis -lact series
Blood group glycolipid biosynthesis -neolact series
Globoside metabolism
Ganglioside biosynthesis
Ceramide-sulfate metabolism
3.2.1.18
2.8.2.11
Sulfatide
3.1.6.1
3.1.6.8
3.2.1.22
2.4.1.-
Digalactosylceramide
2.8.2.11
Digalactosylceramidesulfate
00600 1/26/05

Display



Score, p-value, for TSG or Oncogene

Gene Expression

Controls:
- Threshold
- Patients,
- Percentiles, etc

Preliminary Analysis

- 22 32530895 33077888; pvalue: 1.310680e-01
  - LARGE like-glycosyltransferase
  - GOOD CANDIDATE ?
  - "The function of this gene has not yet been established; however, it may involve a role in tumor-specific genomic rearrangements. Mutations in this gene may be involved in the development and progression of meningioma through modification of ganglioside composition and other glycosylated molecules in tumor cells."
  - http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=retrieve&dopt=full_report&list_uids=9215

Preliminary Analysis

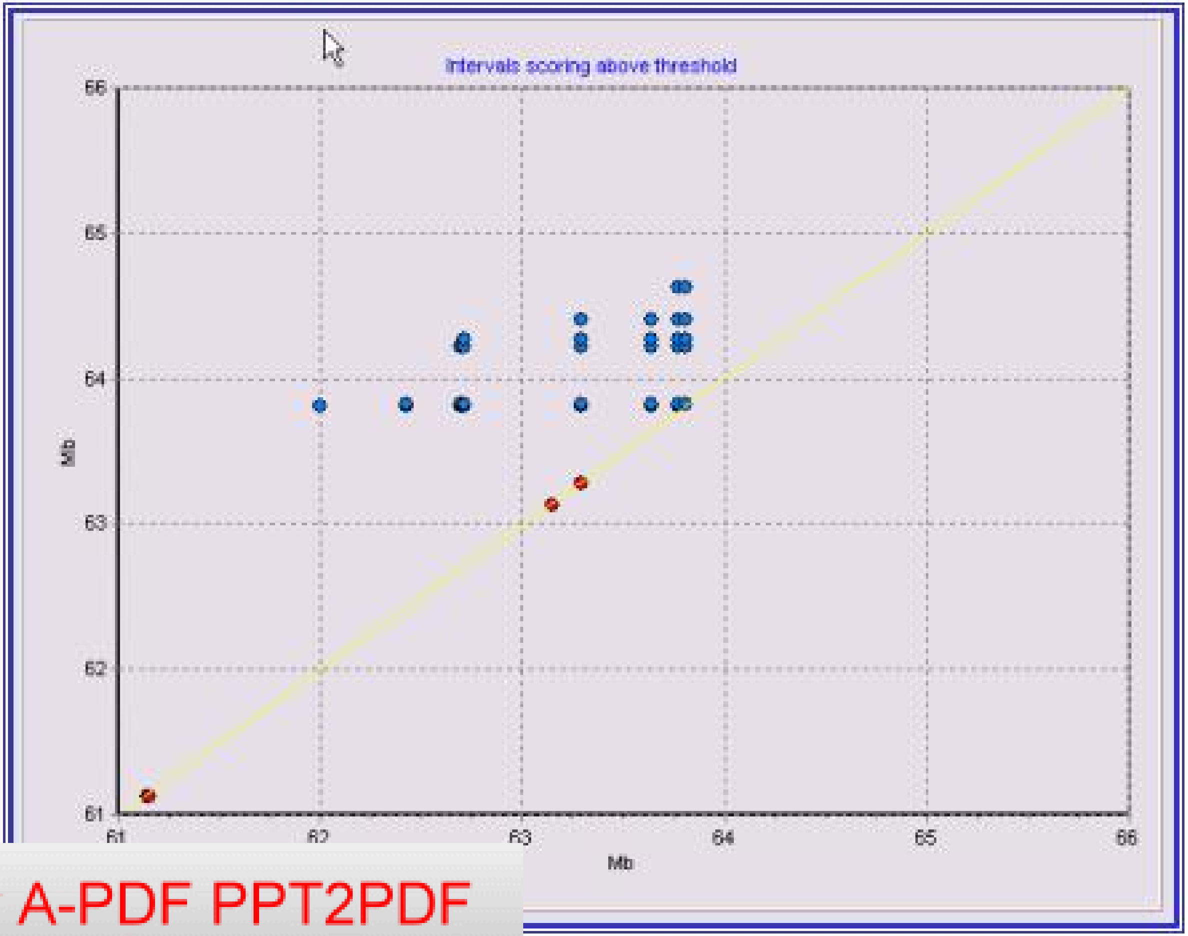
- 9 5890426 6019644; pvalue: 4.344025e-44
  - MLANA melan-A
  - role is unclear
  - Entrez gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=retrieve&dopt=full_report&list_uids=2315
    - Paper 1:"Newly established clear cell sarcoma (malignant melanoma of soft parts) cell line expressing melanoma-associated Melan-A antigen and overexpressing C-MYC oncogene."--? Melan-A is expressed in sarcoma ?
    - http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed&cmd=Retrieve&list_uids=12072203
    - Paper 2:"The lack of melanoma-associated antigen (MAA) expression has been associated with the reduced overall survival in melanoma patients"
    - -- lower expression --> lower survival rate
    - http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed&cmd=Retrieve&list_uids=15305155

Preliminary Analysis

- 15 63803230 63817520;pvalue: 1.433303e-33
  - DENND4A DENN/MADD domain containing 4A
  - "c-myc promoter binding protein"
  - http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=retrieve&dopt=full_report&list_uids=10260
  - interacts with TP73, tumor protein 73
  - http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=Retrieve&dopt=full_report&list_uids=7161

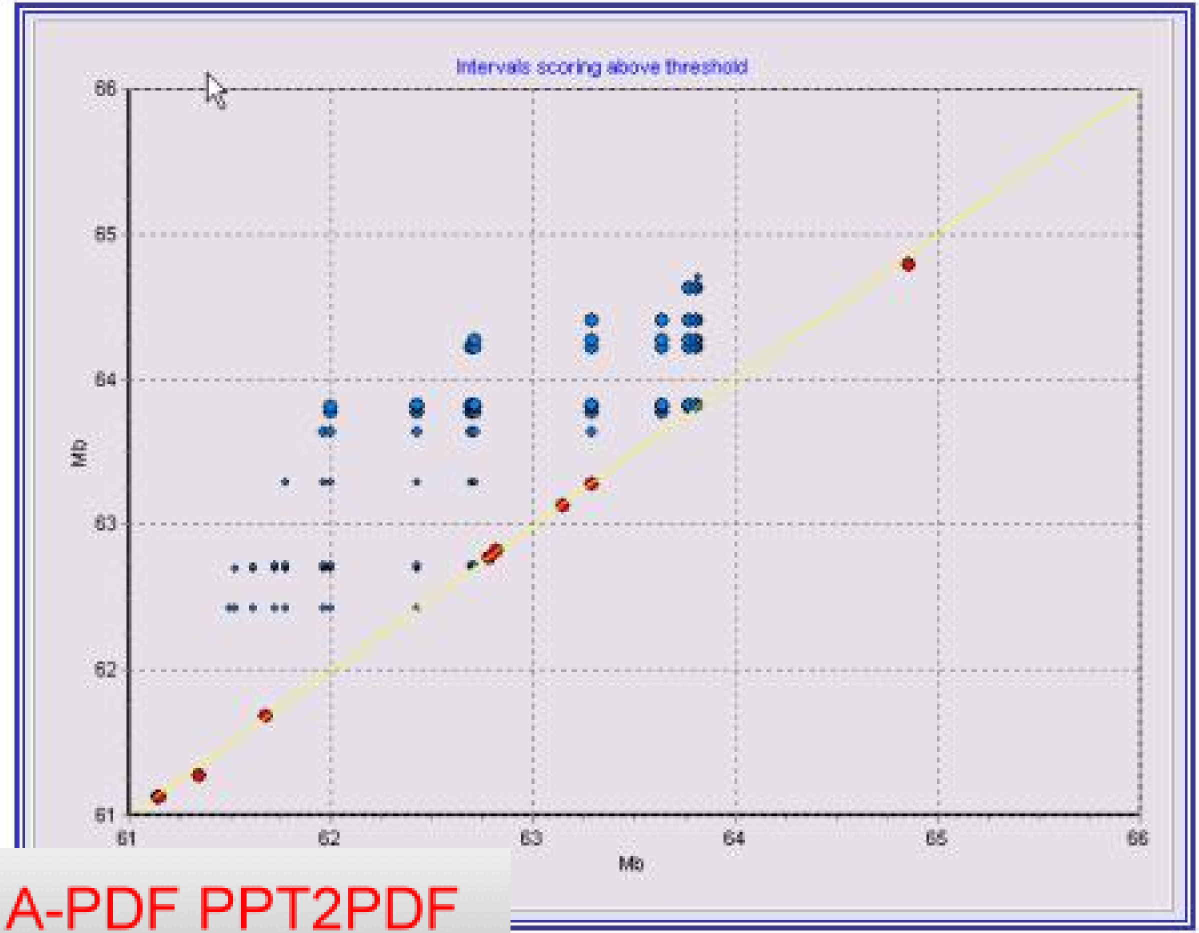
Chromosme 15, ~63Mb

Example



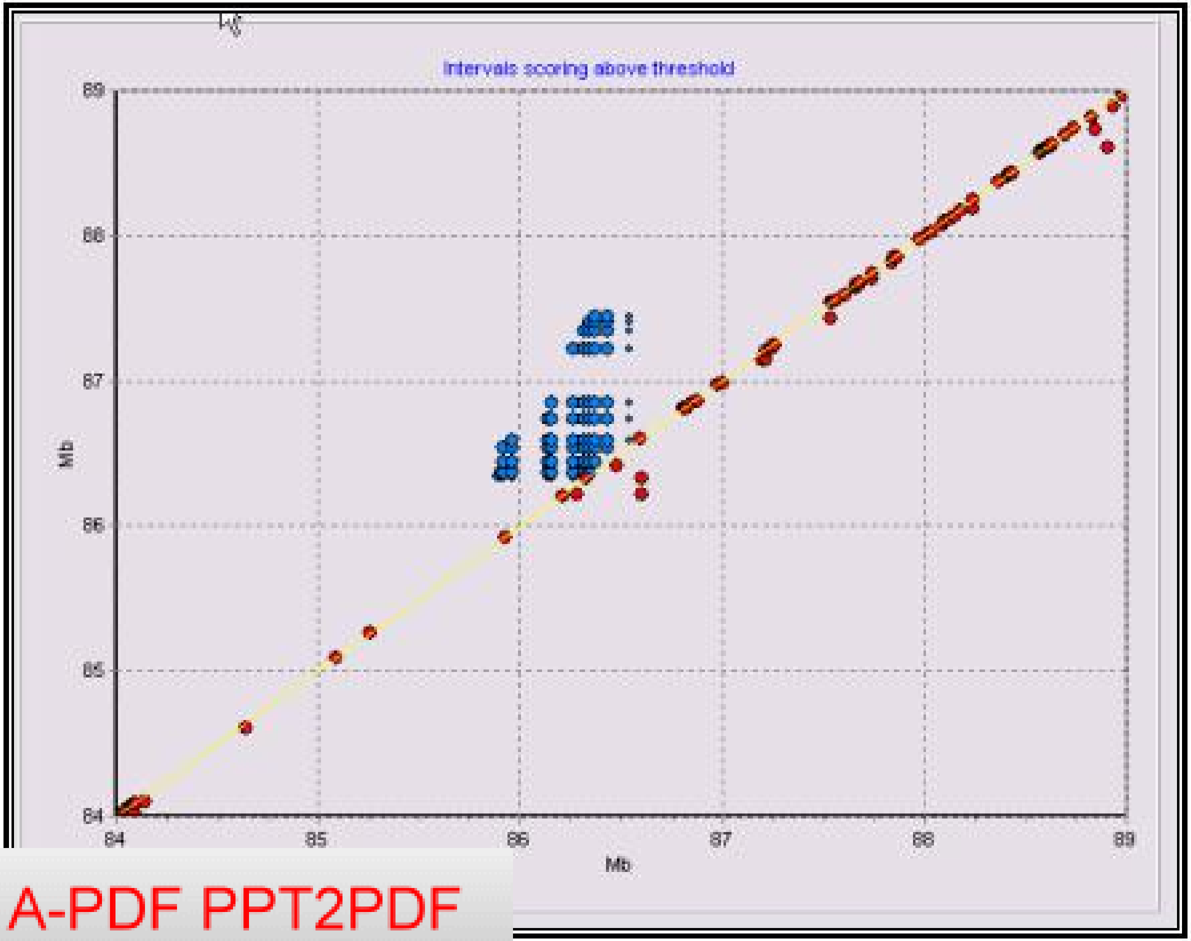
Chromosme 15, ~63Mb

Example



Chromosme 15, ~86Mb

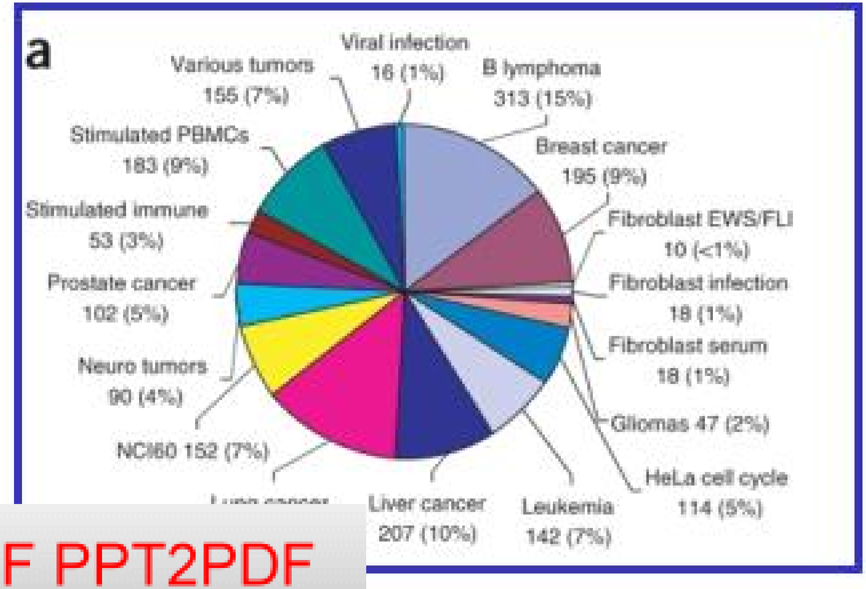
Example



Global Maps

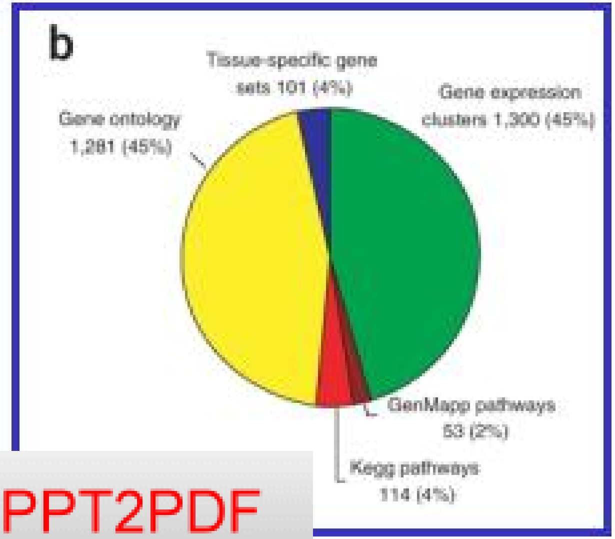
Data

- 14,145 genes
- 1,975 microarrays
- spanning 17 categories (according to the condition they present)

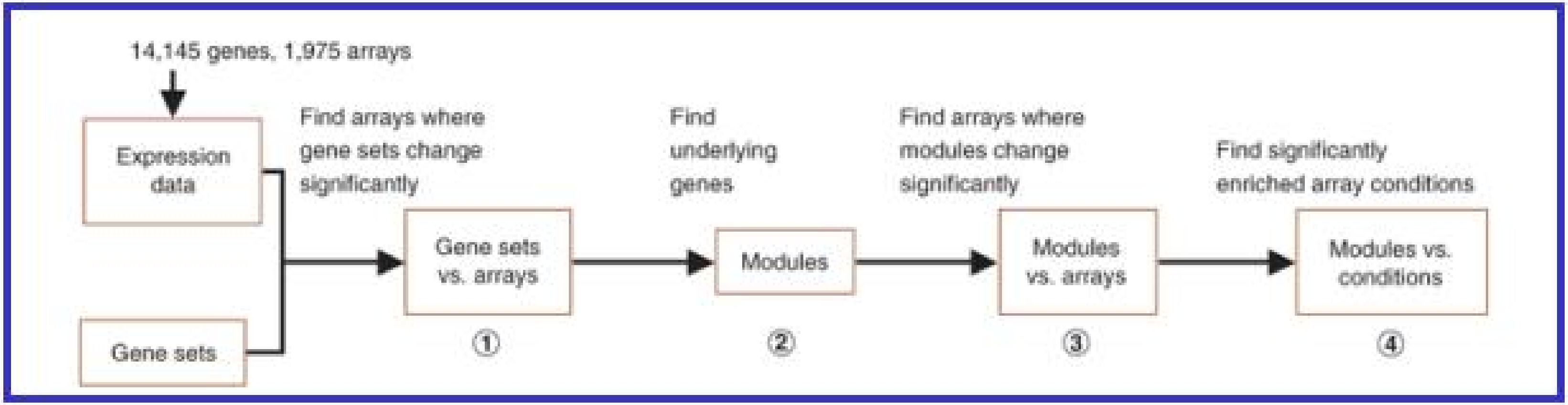


Data (contd.)

- 2,849 biologically meaningful gene sets
  - coexpressed genes
  - genes expressed in specific tissue types
  - genes belonging to the same functional category or pathway

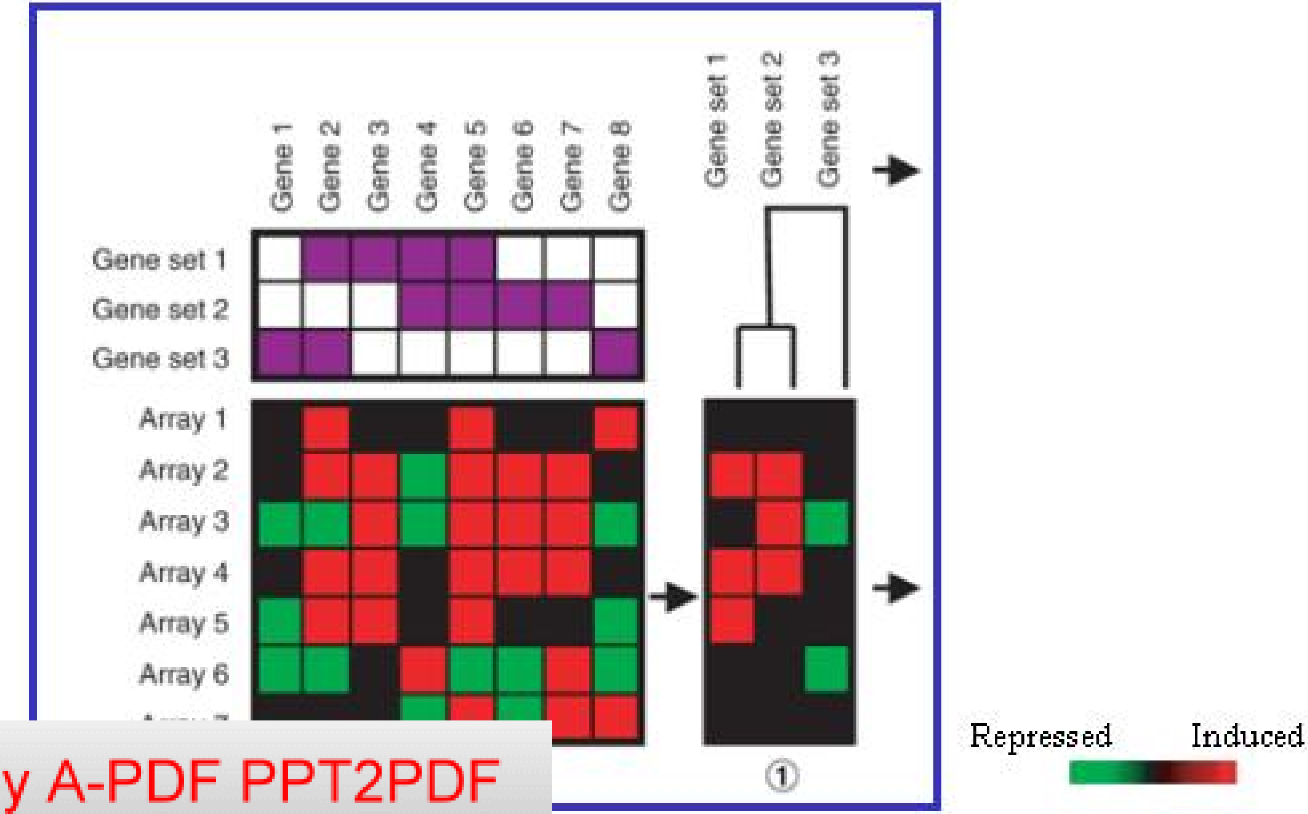


Computational Pipeline


14,145 genes, 1,975 arrays
Expression data
Gene sets
Find arrays where gene sets change significantly
Gene sets vs. arrays
①
Find underlying genes
Modules
②
Find arrays where modules change significantly
Modules vs. arrays
③
Find significantly enriched array conditions
Modules vs. conditions
④

Step 1

- Find arrays where gene sets change significantly

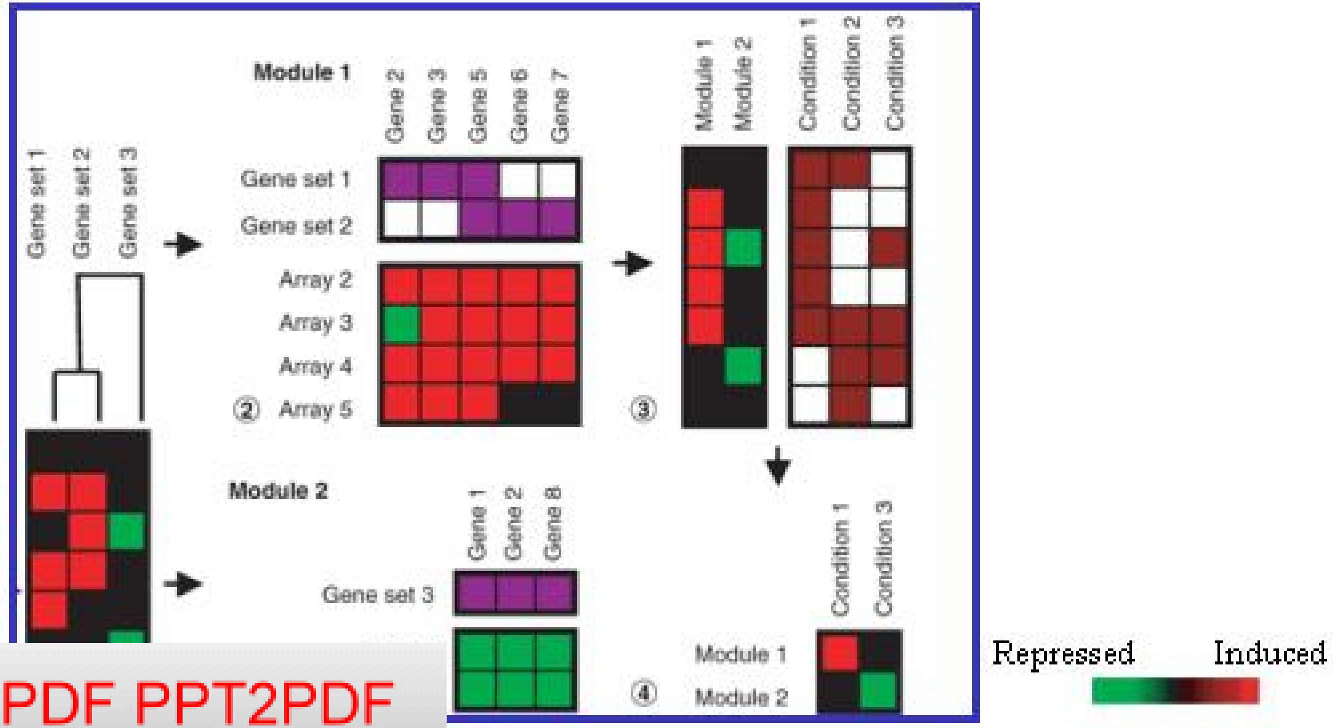


From Gene Sets to Modules

- Gene sets reflect biological modules only approximately
- Only a subset of genes in a set may contribute to its expression signature
- Different gene sets may have similar signature across the arrays
  - Overlap btw the gene sets
  - coregulation of nonoverlaping gene sets
- Module: combines several related gene sets

Steps 2-4

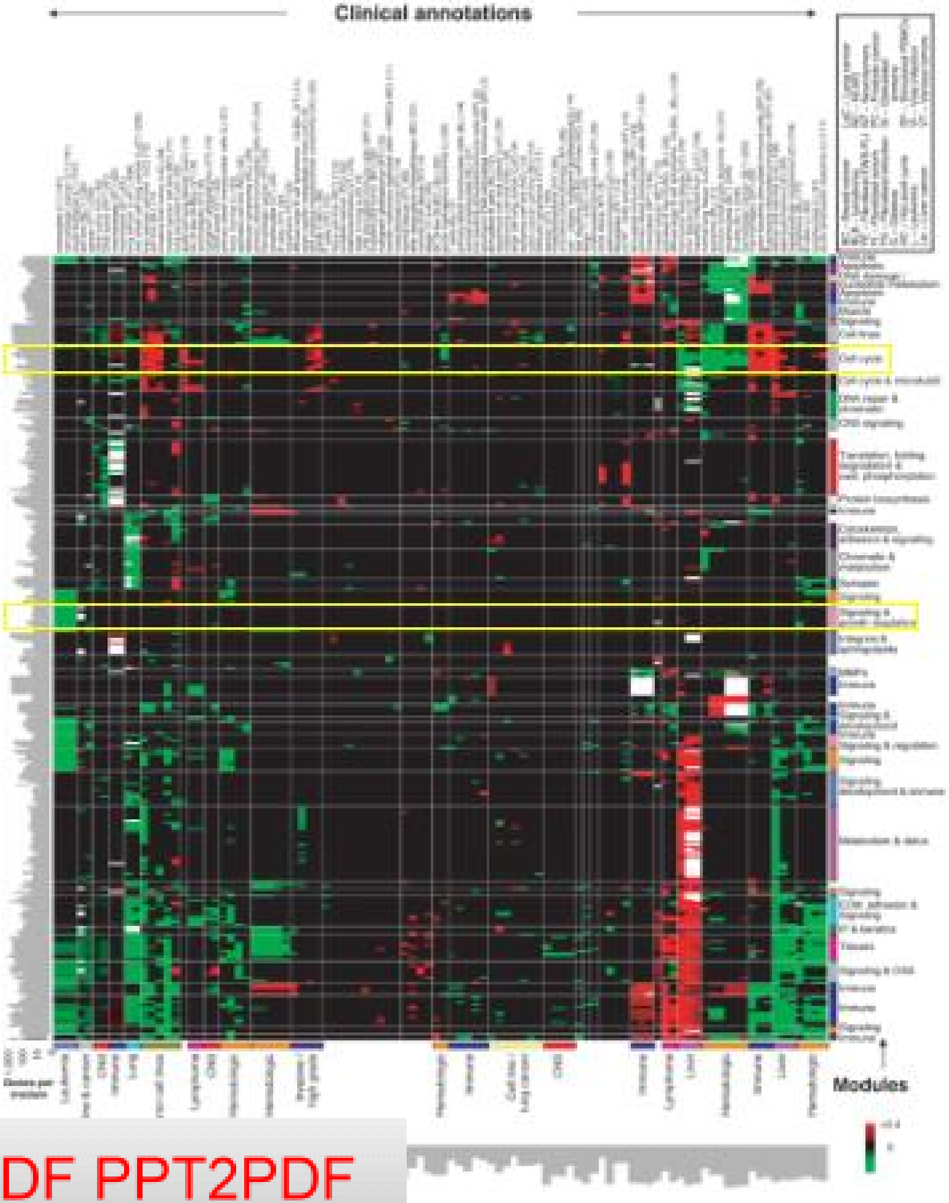
- Find arrays where modules change significantly
- Find significantly enriched array conditions



Modules and Conditions

- 456 statistically significant modules, spanning various processes and functions
  - metabolism, transcription, translation
  - growth, cell cycle, apoptosis
- 263 biological and clinical conditions
  - tissue and tumor type
  - diagnostic and prognostic information

Cancer Module Map



Cell cycle

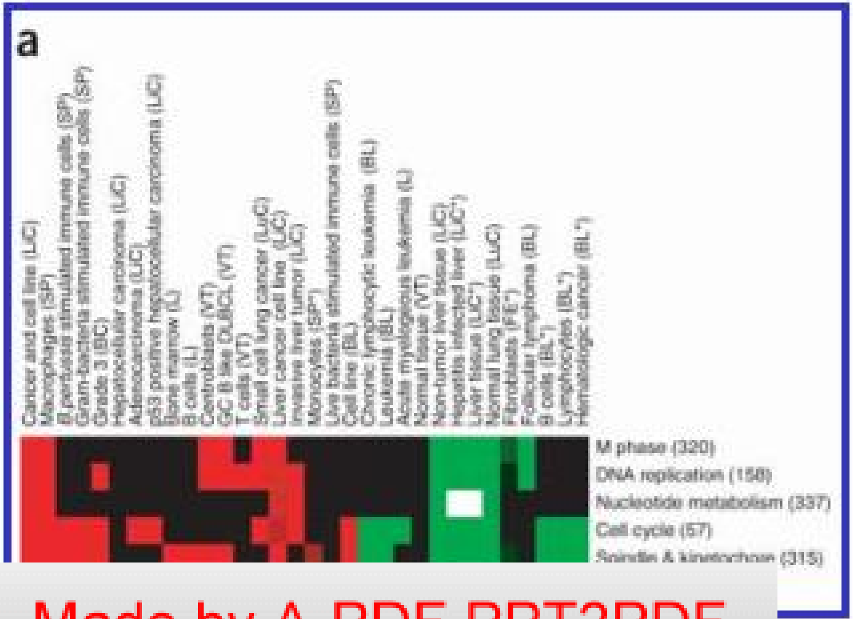
Signaling& growth regulatio n

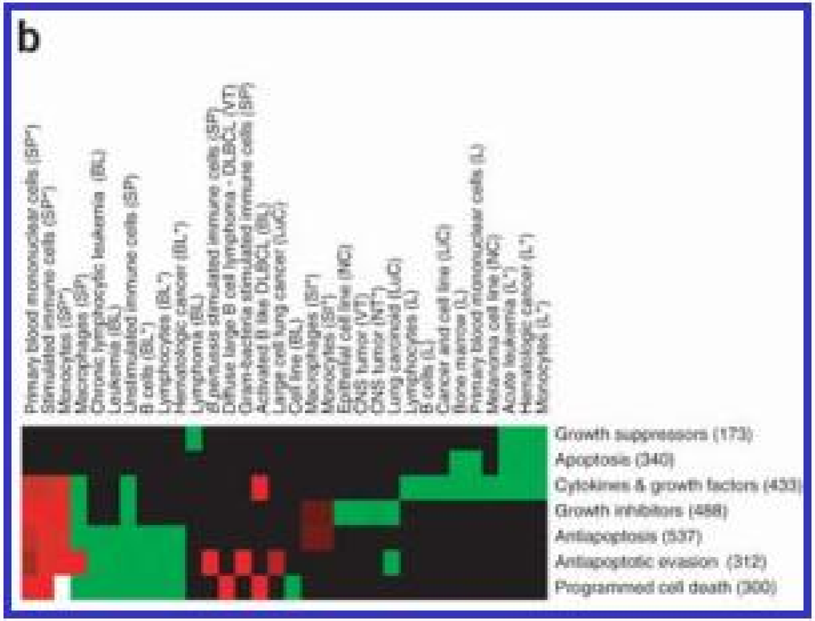
Some Observations

- Some modules are shared across multiple tumor types
  - cell cycle
- Some modules are more specific to the tissue origin or progression of particular tumors
  - modules related to neural processes are repressed in a subset of brain tumors (relative to other central nervous system tumors)

Examples

- Related modules such as cell cycles modules (a)
- More specialized modules such as growth regulatory (b)



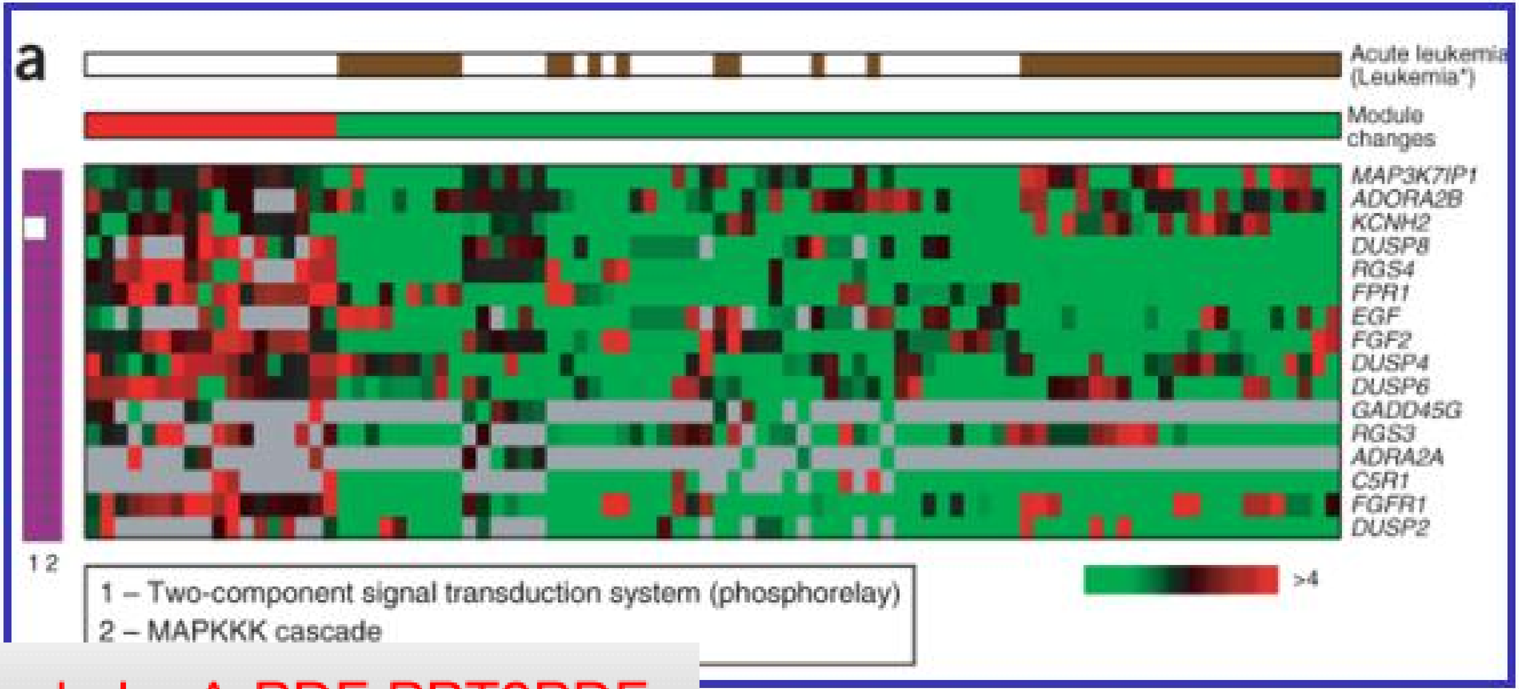


Uses of Module Map

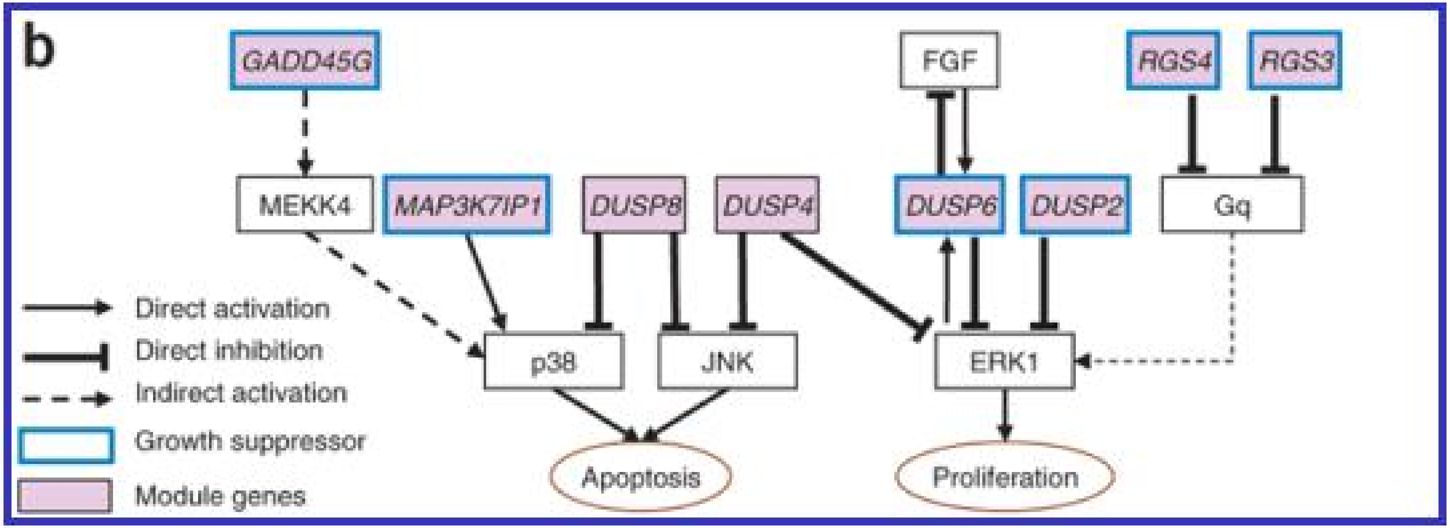
- Characterizes each condition by a particular combination of modules
- Indicates that related conditions involve related modules, although in distinct ways
  - the pattern of involvement separates different tumor types and subtypes
- Gives important insights into the mechanism underlying specific malignancies

Growth Inhibitory Module

- Responds significantly to one specific condition:
  - acute Leukemia

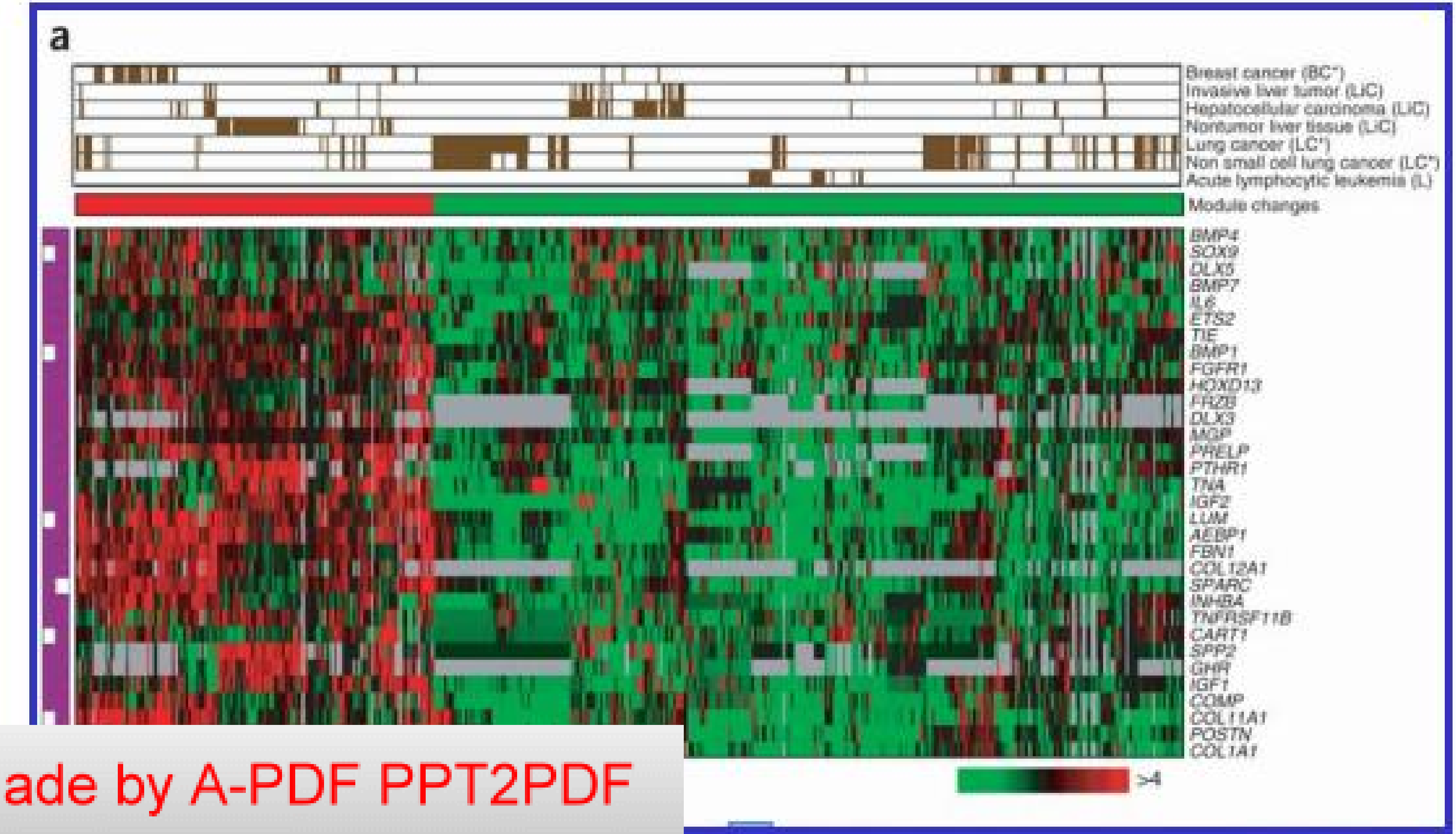


From Module to Pathway



Bone Osteopathic Module

- Responds significantly to multiple conditions:



Take aways

- Provides global view of cancer
- Shows Tumors can be characterized by combinations of a relatively small number of modules
- Uses existing biological knowledge directly, in the form of gene sets and clinical annotations
- Provides a valuable tool for understanding the molecular basis of cancer
  - specific tumors
  - tumorigenic processes in general

Redescription mining

- Our own expressive algorithm (CARTwheels) for relating biological vocabularies [KDD 2004]
- Input
  - Gene set
  - Gene subsets (multiple vocabularies)
- Output
  - Equivalence relationships

To be continued…

• • •