OncoSNP Illumina, MIP, Affy 10K, stromal contamination, and other models

Russell Hanson [2/17/2011]
@Dana Farber Cancer Institute
Boston, MA

OncoSNP Illumina, MIP, Affy 10K, stromal contamination, and other models

OncoSNP was designed for Illumina data

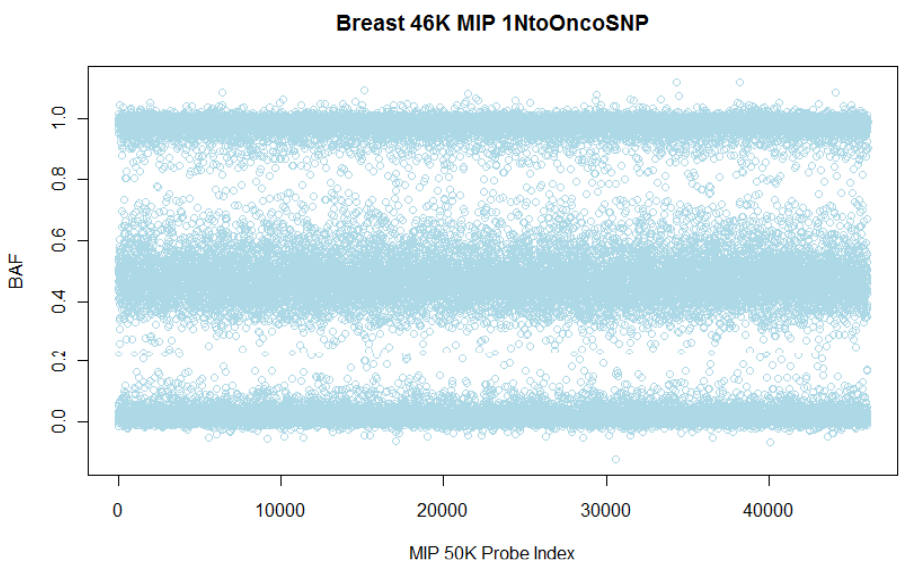
"However, the methods described are not intrinsically tied to the Illumina platform and we are actively working to transfer these techniques for use with the Affmetrix platform."

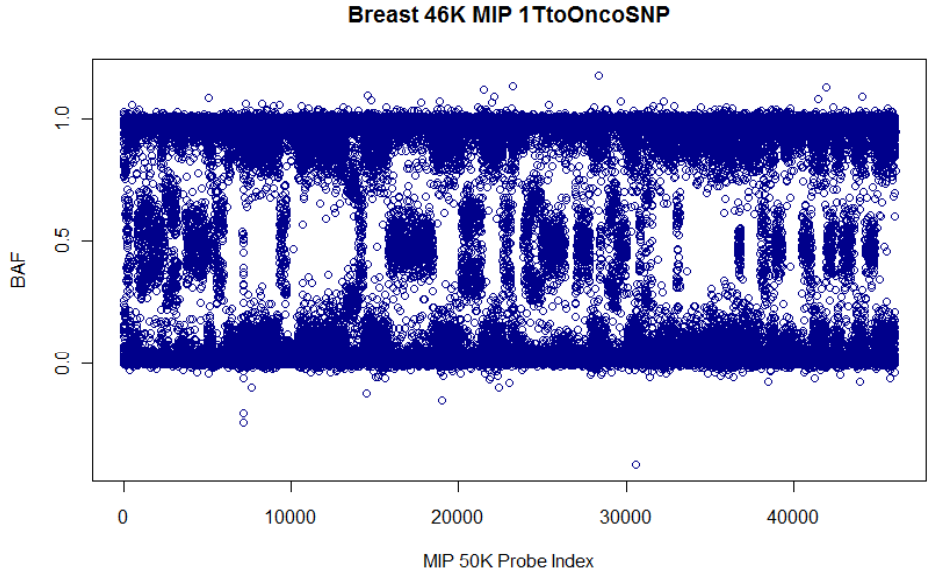
- From "A statistical method for detecting genomic aberrations in heterogeneous tumour samples from single nucleotide polymorphism genotyping data", Yau et al. (2010)

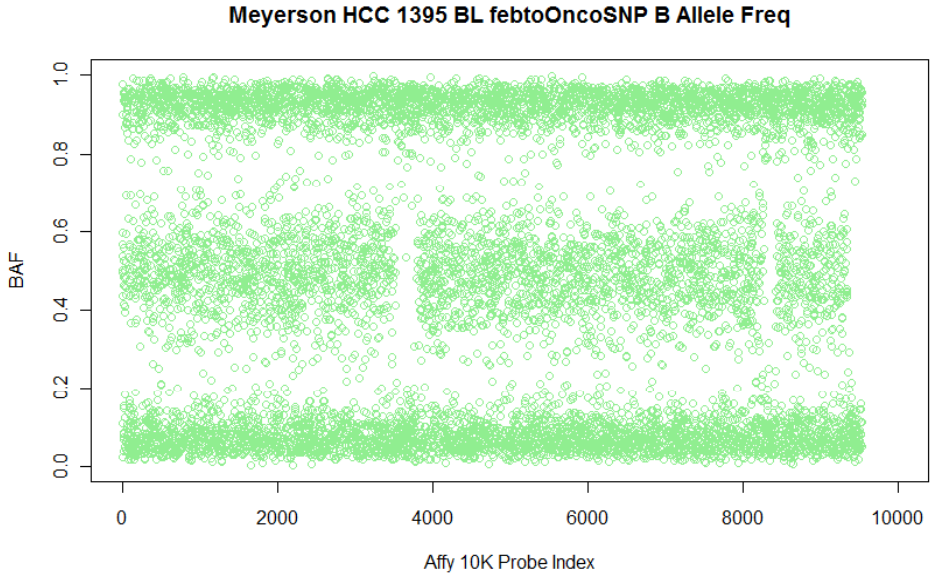
Comparing distribution of B allele frequency between Illumina, MIP and Affymetrix

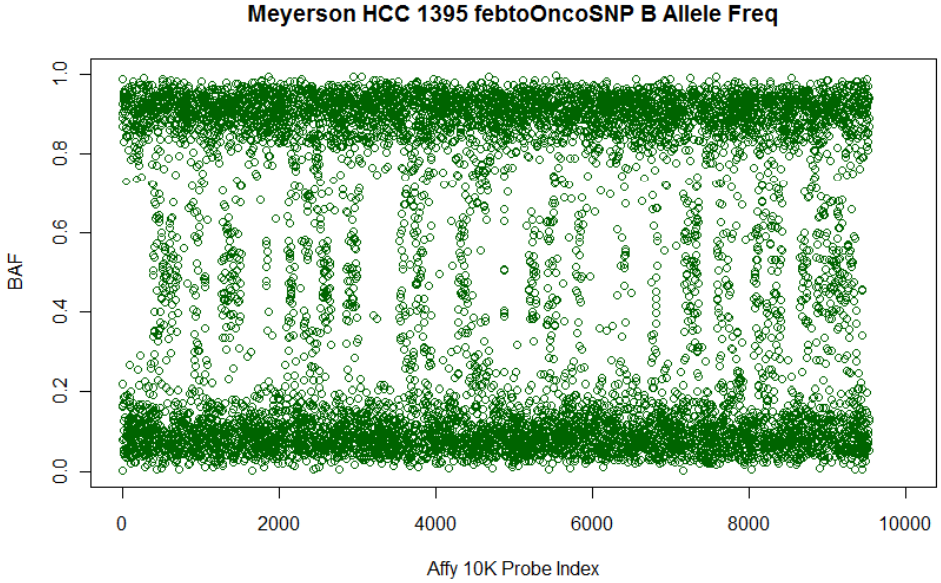
MIP Affy 250K

Affy 10K Illumina

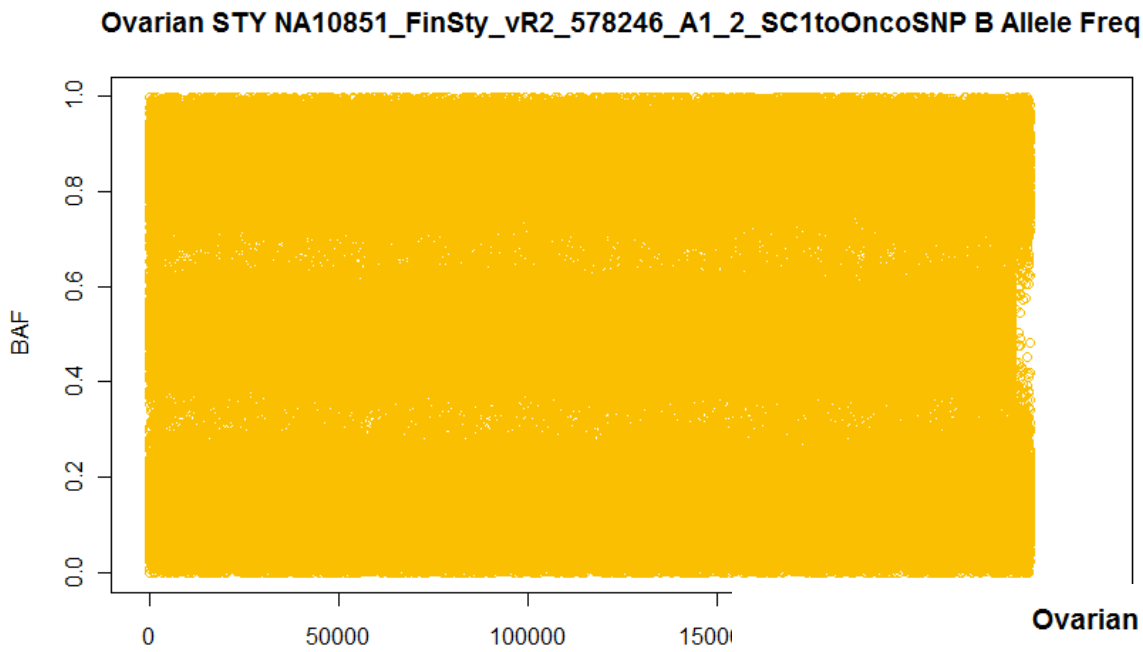


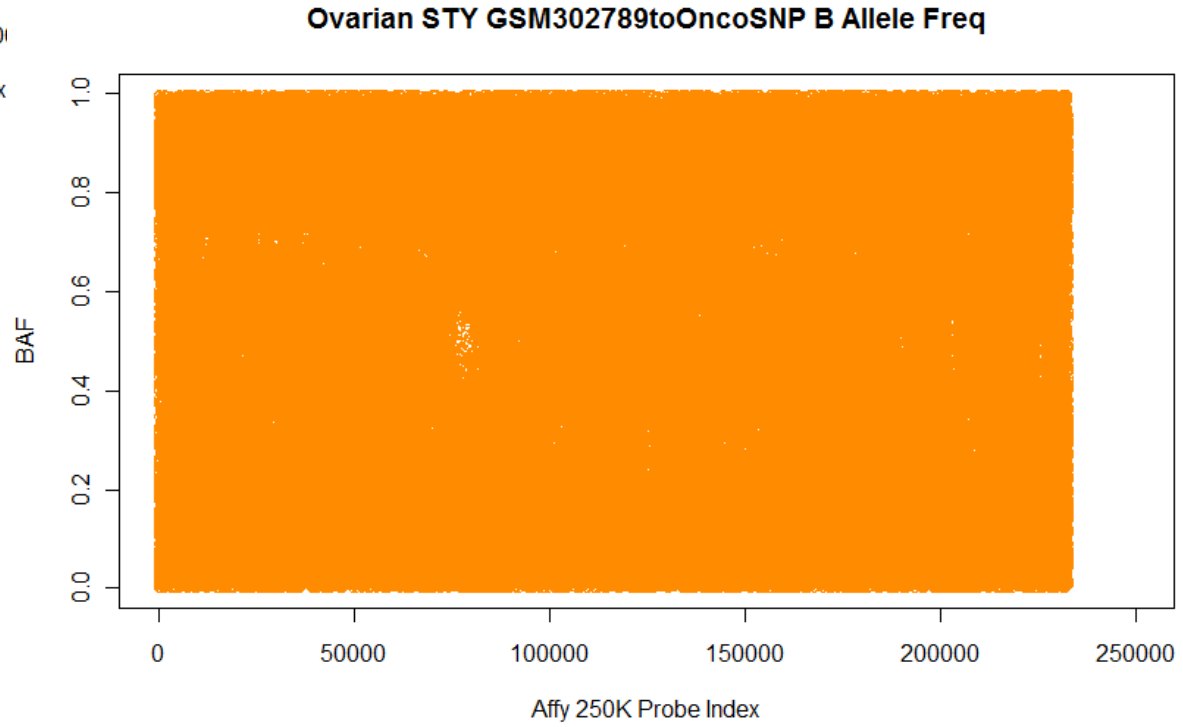




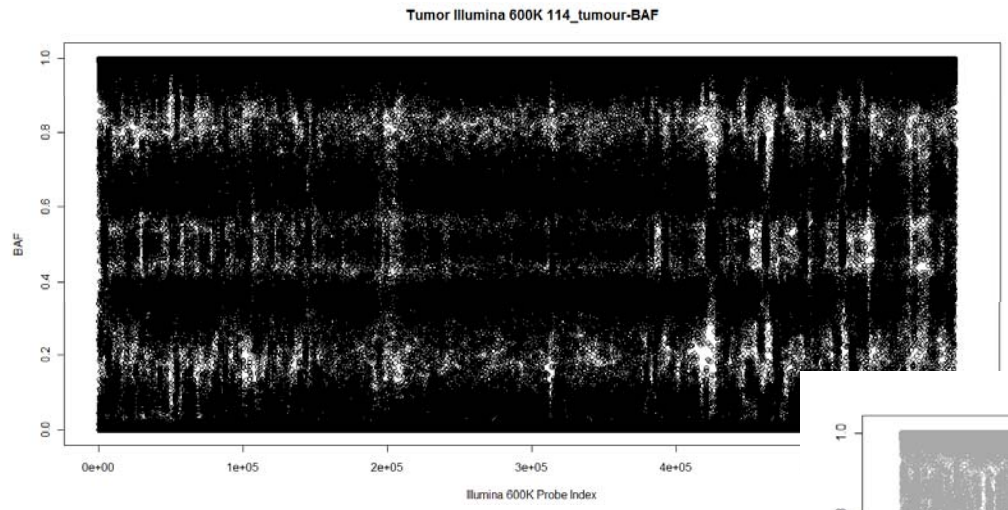


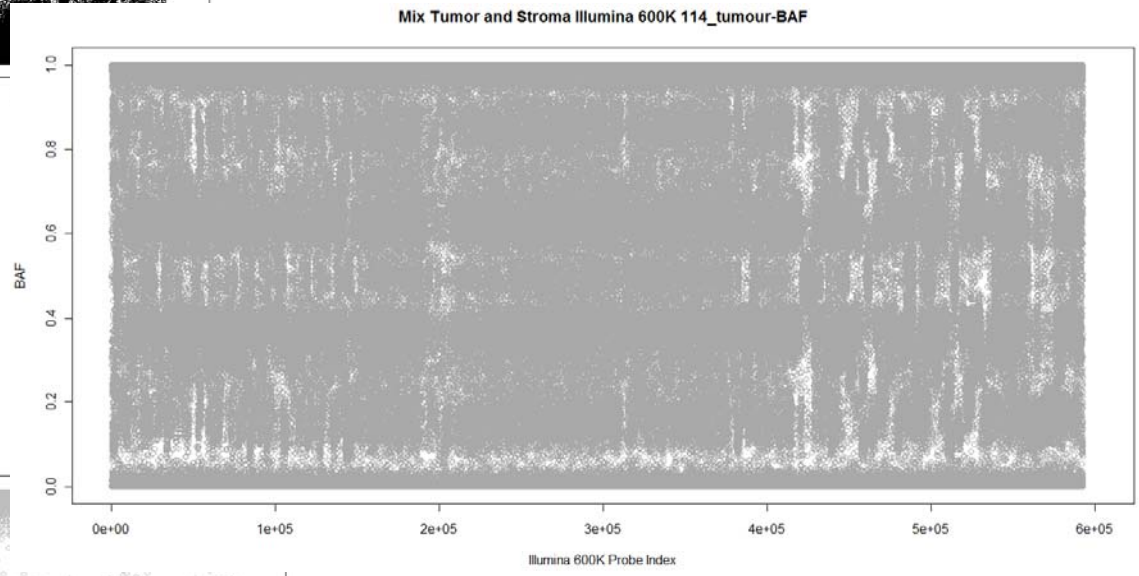
Affymetrix 250K/SNP5 - Comparing distribution of B allele frequency between Illumina, MIP and Affymetrix

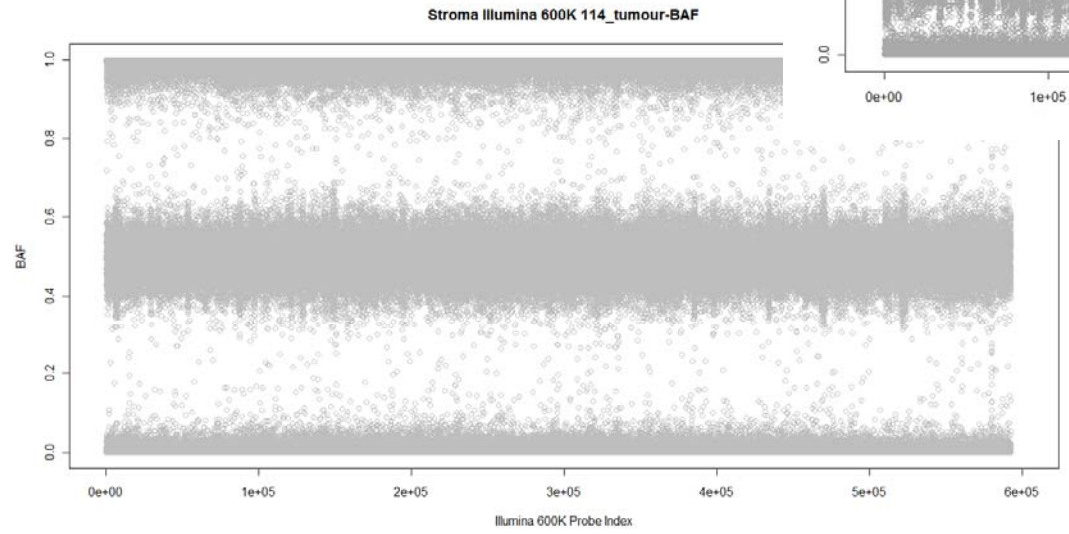




Illumina 600K “Duo” - Comparing distribution of B allele freq between Illumina, MIP and Affymetrix







OncoSNP on Affymetrix 10K data – does not work

SampleID	TumourFile	NormalFile	OncoSNP Output
HCC1187toOncoSNP	HCC 1187toOncoSNP.xls	HCC 1187 BLtoOncoSNP.xls	
HCC1395febtoOncoSNP	HCC 1395 febtoOncoSNP.xls	HCC 1395 BL febtoOncoSNP.xls	X
HCC1008toOncoSNP	HCC1008toOncoSNP.xls	HCC1007 BLtoOncoSNP.xls	X
HCC1143toOncoSNP	HCC1143toOncoSNP.xls	HCC1143 BLtoOncoSNP.xls	
HCC1143M6toOncoSNP	HCC1143M6toOncoSNP.xls	HCC1143 BLtoOncoSNP.xls	X
HCC1143M7toOncoSNP	HCC1143M7toOncoSNP.xls	HCC1143 BLtoOncoSNP.xls	
HCC1143M8toOncoSNP	HCC1143M8toOncoSNP.xls	HCC1143 BLtoOncoSNP.xls	
HCC1143M9toOncoSNP	HCC1143M9toOncoSNP.xls	HCC1143 BLtoOncoSNP.xls	X
HCC1395junetoOncoSNP	HCC1395 junetoOncoSNP.xls	HCC1395BL junetoOncoSNP.xls	
HCC1599toOncoSNP	HCC1599toOncoSNP.xls	HCC1599 BLtoOncoSNP.xls	X
HCC1937toOncoSNP	HCC1937toOncoSNP.xls	HCC1937 BLtoOncoSNP.xls	
HCC2218toOncoSNP	HCC2218toOncoSNP.xls	HCC2218 BLtoOncoSNP.xls	
HCC38marchtoOncoSNP	HCC38 marchtoOncoSNP.xls	HCC38 BL marchtoOncoSNP.xls	X
HCC38maytoOncoSNP	HCC38 maytoOncoSNP.xls	HCC38 BL maytoOncoSNP.xls	
HCC38M6toOncoSNP	HCC38M6toOncoSNP.xls	HCC38 BL maytoOncoSNP.xls	
HCC38M7toOncoSNP	HCC38M7toOncoSNP.xls	HCC38 BL maytoOncoSNP.xls	X
HCC38M8toOncoSNP	HCC38M8toOncoSNP.xls	HCC38 BL maytoOncoSNP.xls	X
HCC38M9toOncoSNP	HCC38M9toOncoSNP.xls	HCC38 BL maytoOncoSNP.xls	X
10372TtoOncoSNP	10372TtoOncoSNP.xls	24149NtoOncoSNP.xls	
18252TtoOncoSNP	18252TtoOncoSNP.xls	60596NtoOncoSNP.xls	X
57588TtoOncoSNP	57588TtoOncoSNP.xls	73315NtoOncoSNP.xls	
83437TtoOncoSNP	83437TtoOncoSNP.xls	73315NtoOncoSNP.xls	
H128toOncoSNP	H128toOncoSNP.xls	HCC1937 BLtoOncoSNP.xls	
H1395toOncoSNP	H1395toOncoSNP.xls	HCC1937 BLtoOncoSNP.xls	
H1648toOncoSNP	H1648toOncoSNP.xls	HCC1937 BLtoOncoSNP.xls	
H2107toOncoSNP	H2107toOncoSNP.xls	HCC1937 BLtoOncoSNP.xls	
H2141toOncoSNP	H2141toOncoSNP.xls	HCC1937 BLtoOncoSNP.xls	
H2171toOncoSNP	H2171toOncoSNP.xls	HCC1937 BLtoOncoSNP.xls	X
H289toOncoSNP	H289toOncoSNP.xls	HCC1937 BLtoOncoSNP.xls	X
BT474toOncoSNP	BT474toOncoSNP.xls	HCC1937 BLtoOncoSNP.xls	X
MCF7toOncoSNP	MCF7toOncoSNP.xls	HCC1937 BLtoOncoSNP.xls	X
NA01201BtoOncoSNP	NA01201BtoOncoSNP.xls	HCC1937 BLtoOncoSNP.xls	X
NA01416toOncoSNP	NA01416toOncoSNP.xls	HCC1937 BLtoOncoSNP.xls	
NA01723toOncoSNP	NA01723toOncoSNP.xls	HCC1937 BLtoOncoSNP.xls	
NA03226toOncoSNP	NA03226toOncoSNP.xls	HCC1937 BLtoOncoSNP.xls	X
NA04626toOncoSNP	NA04626toOncoSNP.xls	HCC1937 BLtoOncoSNP.xls	
NA06061toOncoSNP	NA06061toOncoSNP.xls	HCC1937 BLtoOncoSNP.xls	X
UACC812toOncoSNP	UACC812toOncoSNP.xls	HCC1937 BLtoOncoSNP.xls	

Total: 17/38

SampleID	TumourFile	NormalFile	OncoSNP Output
T115toOncoSNP	T115toOncoSNP.xls	B115toOncoSNP.xls	
T116toOncoSNP	T116toOncoSNP.xls	B116toOncoSNP.xls	X
T117toOncoSNP	T117toOncoSNP.xls	B117toOncoSNP.xls	X
T118toOncoSNP	T118toOncoSNP.xls	B118toOncoSNP.xls	
T119toOncoSNP	T119toOncoSNP.xls	B119toOncoSNP.xls	X
T123toOncoSNP	T123toOncoSNP.xls	B123toOncoSNP.xls	X
T125toOncoSNP	T125toOncoSNP.xls	B118toOncoSNP.xls	X
T127toOncoSNP	T127toOncoSNP.xls	B118toOncoSNP.xls	X
T129toOncoSNP	T129toOncoSNP.xls	B129toOncoSNP.xls	
T130toOncoSNP	T130toOncoSNP.xls	B130toOncoSNP.xls	
T133toOncoSNP	T133toOncoSNP.xls	B133toOncoSNP.xls	
T134toOncoSNP	T134toOncoSNP.xls	B134toOncoSNP.xls	X
T137toOncoSNP	T137toOncoSNP.xls	B137toOncoSNP.xls	
T140toOncoSNP	T140toOncoSNP.xls	B140toOncoSNP.xls	X
T141toOncoSNP	T141toOncoSNP.xls	B141toOncoSNP.xls	X
T143toOncoSNP	T143toOncoSNP.xls	B143toOncoSNP.xls	
T144toOncoSNP	T144toOncoSNP.xls	B144toOncoSNP.xls	
T145toOncoSNP	T145toOncoSNP.xls	B145toOncoSNP.xls	X
T146toOncoSNP	T146toOncoSNP.xls	B146toOncoSNP.xls	X
T147toOncoSNP	T147toOncoSNP.xls	B118toOncoSNP.xls	X
T149toOncoSNP	T149toOncoSNP.xls	B149toOncoSNP.xls	
T151toOncoSNP	T151toOncoSNP.xls	B151toOncoSNP.xls	X
T152toOncoSNP	T152toOncoSNP.xls	B118toOncoSNP.xls	X
T161toOncoSNP	T161toOncoSNP.xls	B161toOncoSNP.xls	
T162toOncoSNP	T162toOncoSNP.xls	B118toOncoSNP.xls	X
T173toOncoSNP	T173toOncoSNP.xls	B118toOncoSNP.xls	X
T175F3toOncoSNP	T175F3toOncoSNP.xls	B175toOncoSNP.xls	
T178toOncoSNP	T178toOncoSNP.xls	B118toOncoSNP.xls	X
T183toOncoSNP	T183toOncoSNP.xls	B183toOncoSNP.xls	X
T21toOncoSNP	T21toOncoSNP.xls	B21toOncoSNP.xls	X
T27toOncoSNP	T27toOncoSNP.xls	B118toOncoSNP.xls	X
T30toOncoSNP	T30toOncoSNP.xls	B30toOncoSNP.xls	
T37toOncoSNP	T37toOncoSNP.xls	B37toOncoSNP.xls	
T38toOncoSNP	T38toOncoSNP.xls	B38toOncoSNP.xls	X
T41toOncoSNP	T41toOncoSNP.xls	B41toOncoSNP.xls	X
T44toOncoSNP	T44toOncoSNP.xls	B44toOncoSNP.xls	
T45toOncoSNP	T45toOncoSNP.xls	B45toOncoSNP.xls	X
T46toOncoSNP	T46toOncoSNP.xls	B118toOncoSNP.xls	X
T4toOncoSNP	T4toOncoSNP.xls	B4toOncoSNP.xls	X
T50toOncoSNP	T50toOncoSNP.xls	B50toOncoSNP.xls	X
T56toOncoSNP	T56toOncoSNP.xls	B56toOncoSNP.xls	X
T636toOncoSNP	T636toOncoSNP.xls	B118toOncoSNP.xls	X
T72toOncoSNP	T72toOncoSNP.xls	B118toOncoSNP.xls	
T73toOncoSNP	T73toOncoSNP.xls	B73toOncoSNP.xls	X
T74toOncoSNP	T74toOncoSNP.xls	B118toOncoSNP.xls	
T80toOncoSNP	T80toOncoSNP.xls	B80toOncoSNP.xls	X
T81toOncoSNP	T81toOncoSNP.xls	B81toOncoSNP.xls	
T84toOncoSNP	T84toOncoSNP.xls	B84toOncoSNP.xls	X
T911toOncoSNP	T911toOncoSNP.xls	B911toOncoSNP.xls	
T92toOncoSNP	T92toOncoSNP.xls	B92toOncoSNP.xls	X

Total: 32/50

OncoSNP on MIP 46K – comparable to dChip; OncoSNP breaks down with high contamination -- not built for this platform

1	Tumor	OncoSNP-dChip LOH correlation run1	Previously removed b/c contamination	run2
2	1T	0.9120419		0.912041
3	3T	0.8915712		0.891571
4	4T	0.2428070		0.242806
5	5T	0.8671010		0.867101
6	6NP_FFPE	NA		NA
7	6T	0.9408540		0.940854
8	7NP_FFPE	NA		NA
9	7T	0.9598840		0.959884
10	8NP_FFPE	NA		NA
11	8T	0.8952590		0.895259
12	9T_FFPE	0.8602720		0.860272
13	10T	0.2182451	X	0.218245
14	10UE	0.0388197		0.002697
15	11T	0.0921312		0.092131
16	11UE	NA		NA
17	12T	0.3689727		0.368972
18	13T_FFPE	0.9344420		0.934442
19	14T	0.8865220		0.886522
20	15T	-0.0060460	X	-0.006046
21	15UE	0.0541220		0.04157
22	16T	0.5476580		0.547658
23	17T	-0.0132228	X	-0.013222
24	18T_FFPE	-0.0579694	X	-0.057969
25	20T	0.0857630	X	0.085763
26	21T	0.0759889		0.075988
27	22T	0.7834360		0.783436
28	23T	0.8975019		0.897501
29	24T	0.2471804		0.24718
30	25T	-0.0022492		-0.002249
31	26T	0.2113694		0.211369
32	27T	0.5496994		0.549699
33	28T	-0.0579091	X	-0.057909
34	28UE	NA		NA
35	29T	0.6917749		0.691774

How MCP is calculated in dChip

Major copy proportion (MCP) of a SNP is defined as C_2/(C_1 + C_2) where C_1 and C_2 are the parental copy numbers at this SNP in a sample and C_1 <= C_2. Total copy number is defined as C_1 + C_2.

MCP is 0.5 for normal loci or balanced copy number alterations, 1 for LOH, and an intermediate value between 0.5 and 1 for allelic imbalanced copy number alterations.

MCP values to be inferred using the HMM are in the range 0.5 to 1 and have 11 states under the default increasing step of 0.05 (comparable to the noise level in the data). The Viterbi algorithm is used to obtain the most probable MCP state path as the inferred MCP values.

A composite alteration score using both MCP and total copy number may also be used, such as the proportion of samples with copy > 3 and MCP > 0.65 to capture only allelic imbalanced amplifications.

Quite a number of allele-specific tumor and stromal CNV studies/algorithms…

1. Title: A new analysis tool for individual-level allele frequency for genomic studies
Author(s): Yang HC, Lin HC, Huang MC, et al.
Source: **BMC GENOMICS** Volume: **11** Article Number: **415** Published: **JUL 5 2010**
Times Cited: **0**
Get this – MIT SFX

2. Title: Computational Analysis of Whole-Genome Differential Allelic Expression Data in Human
Author(s): Wagner JR, Ge B, Pokholok D, et al.
Source: **PLOS COMPUTATIONAL BIOLOGY** Volume: **6** Issue: **7** Article Number: **e1000849** Published: **JUL 2010**
Times Cited: **0**
Get this – MIT SFX

3. Title: TumorBoost: Normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays
Author(s): Bengtsson H, Neuvial P, Speed TP
Source: **BMC BIOINFORMATICS** Volume: **11** Article Number: **245** Published: **MAY 12 2010**
Times Cited: **0**
Get this – MIT SFX

4. Title: MixHMM: Inferring Copy Number Variation and Allelic Imbalance Using SNP Arrays and Tumor Samples Mixed with Stromal Cells
Author(s): Liu ZZ, Li A, Schulz V, et al.
Source: **PLOS ONE** Volume: **5** Issue: **6** Article Number: **e10909** Published: **JUN 1 2010**
Times Cited: **0**
Get this – MIT SFX

5. Title: Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays
Author(s): Popova T, Manie E, Stoppa-Lyonnet D, et al.
Source: **GENOME BIOLOGY** Volume: **10** Issue: **11** Article Number: **R128** Published: **2009**
Times Cited: **0**
Get this – MIT SFX

6. Title: PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data
Author(s): Greenman CD, Bignell G, Butler A, et al.
Source: **BIOSTATISTICS** Volume: **11** Issue: **1** Pages: **164-175** Published: **JAN 2010**
Times Cited: **7**
Get this – MIT SFX

7. Title: Two-Round Coamplification at Lower Denaturation Temperature-PCR (COLD-PCR)-Based Sanger Sequencing Identifies a Novel Spectrum of Low-Level Mutations in Lung Adenocarcinoma
Author(s): Li J, Milbury CA, Li C, et al.
Source: **HUMAN MUTATION** Volume: **30** Issue: **11** Pages: **1583-1590** Published: **NOV 2009**
Times Cited: **1**
Get this – MIT SFX

8. Title: Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances
Author(s): LaFramboise T
Source: **NUCLEIC ACIDS RESEARCH** Volume: **37** Issue: **13** Pages: **4181-4193** Published: **JUL 2009**
Times Cited: **14**
Get this – MIT SFX

9. Title: Functional Genomic Analysis Identified Epidermal Growth Factor Receptor Activation as the Most Common Genetic Event in Oral Squamous Cell Carcinoma
Author(s): Sheu JJC, Hua CH, Wan L, et al.
Source: **CANCER RESEARCH** Volume: **69** Issue: **6** Pages: **2568-2576** Published: **MAR 15 2009**
Times Cited: **16**
Get this – MIT SFX

How is LOH Calculated in OncoSNP

- Expectation maximization (EM) 15 iterations

Let π_0 denote the normal DNA fraction of the tumor sample due to stromal contamination and $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^n$ denote the proportion of tumor cells having the normal genotype at each probe. The data $\boldsymbol{y} = \{\boldsymbol{y}_i\}_{i=1}^n$ consists of a set of two-dimensional vectors $\boldsymbol{y}_i = [r_i, b_i]'$ whose elements correspond to the Log R Ratio and B allele frequency respectively.

Given $(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\pi}, \pi_0)$ the data is assumed to be distributed according to a (K + 1)-component mixture of Student t-distributions, where k_i indicates the mixture component assignment of the i-th data point,

$$\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{z}_i, k_i, \boldsymbol{m}, \boldsymbol{\delta}, \boldsymbol{\Sigma} = \begin{cases} \mathrm{St}(\boldsymbol{m}(\boldsymbol{x}_i, \boldsymbol{z}_i) + \boldsymbol{\delta}_{k_i}^{(l_i)}, \sum_{k_i}^{(l_i)}, \nu), & k \neq 0, \\ \mathrm{U}_r(r_{\min}, r_{\max}) \times \mathrm{U}_b(0,1), & k = 0, \end{cases} \quad (1)$$

sentation. These probe measurements called the Log R Ratio and the B Allele Frequency respectively and are defined (approximately) as follows:

$$R = X + Y, \quad (1)$$

$$r = \log(R/R_{ref}), \quad (2)$$

$$b = \frac{Y}{X+Y} + b_{ref}, \quad (3)$$

where r_{ref} and b_{ref} are constants that adjust for probe-specific biases.

Use EM to find a combination of pi and theta with a maximum likelihood where pi is the value for the *stromal contamination*

3 Posterior Inference

Conditional on a value for the stromal contamination π_0, we compute maximum *a posteriori* (MAP) estimates for the model parameters $\theta = \{\eta, \boldsymbol{w}, \boldsymbol{\delta}, \boldsymbol{\Sigma}, \boldsymbol{\beta}\}$ using a expectation-conditional maximisation (ECM) algorithm. We apply the ECM algorithm for a discrete set of values of π_0 between 0 and 1 and find the combination (π_0, θ) that has maximum likelihood.

As the mixture model involves Student t-distributions, we utilise the representation of the Student t-distribution as a scale mixtures of Normal distributions, treating the scaling variables as latent variables in the ECM algorithm,

$$\mathrm{St}(\boldsymbol{y}; \boldsymbol{m}, \boldsymbol{\Sigma}, \nu) = \int_0^\infty \mathrm{N}(\boldsymbol{y}; \boldsymbol{m}, u\boldsymbol{\Sigma})\mathrm{IG}(u; \nu/2, \nu/2)du \tag{25}$$

where $\mathrm{St}(\boldsymbol{y}; \boldsymbol{m}, \boldsymbol{\Sigma}, \nu)$ is the probability density function of the Student t-distribution with mean $\boldsymbol{m}$, covariance $\boldsymbol{\Sigma}$ and ν degrees of freedom, $\mathrm{N}(\boldsymbol{y}; \boldsymbol{m}, \boldsymbol{\Sigma})$ is the probability density function of the Normal distribution with mean $\boldsymbol{m}$ and covariance $\boldsymbol{\Sigma}$ and $\mathrm{IG}(\cdot)$ is the probability density function of the inverse-Gamma distribution with parameters $(\nu/2, \nu/2)$.

The ECM algorithm obtains updated parameter estimates θ' by maximising the expected complete data log-likelihood conditonal on the current estimate $\hat{\theta}$:

$$\theta' = \arg\max_{\theta} \mathbb{E}_{\boldsymbol{x},\boldsymbol{z},\boldsymbol{k},\boldsymbol{u},\boldsymbol{\pi}}[\log p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{k}, \boldsymbol{u}, \boldsymbol{\pi}, \theta)|p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{k}, \boldsymbol{u}, \boldsymbol{\pi}|\boldsymbol{y}, \hat{\theta})], \tag{26}$$

which, under certain regularity conditions, each iteration is guaranteed to increase the likelihood (or posterior probability in this instance).

We can derive a maximum likelihood estimator for the regression coefficients using an expectation-maximisation algorithm which iterates between the following two operations:

$$\boldsymbol{\beta}_j = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y}, \tag{50}$$

$$\sigma^2 = \frac{1}{n} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_j)^T \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_j). \tag{51}$$

The diagonal elements of the $n \times n$ weight matrix $\boldsymbol{W}$ are given by,

$$E(1/V_i | y_i, \boldsymbol{\beta}_j, \sigma^2, \nu) = \frac{1}{\nu\sigma^2 + (y_i - \boldsymbol{X}_i \boldsymbol{\beta}_j)^2}, \; i = 1, \ldots, n. \tag{52}$$

The corrected Log R Ratio value for the j-th probe of the tumour is given by:

$$\tilde{y}_j = y_j - \beta_{j,1} x_j.$$

We use an expectation-maximisation algorithm to learn the HMM parameters. We initialise the transition matrix π to the following,

State, x	Description	$\phi(AA\|x)$	$\phi(AB\|x)$	$\phi(BB\|x)$	$\phi(NC\|x)$
1	Normal	$(1-\nu)/3$	$(1-\nu)/3$	$(1-\nu)/3$	ν
2	Autozygosity	$(1-\nu)/2$	0	$(1-\nu)/2$	ν
3	LOH	$(1-\nu)/2$	0	$(1-\nu)/2$	ν

Table 3: Initial values for the state-conditional emission probabilities.

$$\pi = \begin{pmatrix} 1-\rho_0 & \rho_0/2 & \rho_0/2 \\ \rho_a/2 & 1-\rho_a & \rho_a/2 \\ \rho_l/2 & \rho_l/2 & 1-\rho_l \end{pmatrix} \tag{54}$$

where $(\rho_0 = 0.001, \rho_a = 0.01, \rho_l = 0.001)$ are the transition probabilities out of the normal, autozygosity and LOH states respectively. The state-conditional emission probabilities are initialised to the values given in Table 3 with $\nu = 0.01$.

How is %stromal calculation calculated in OncoSNP

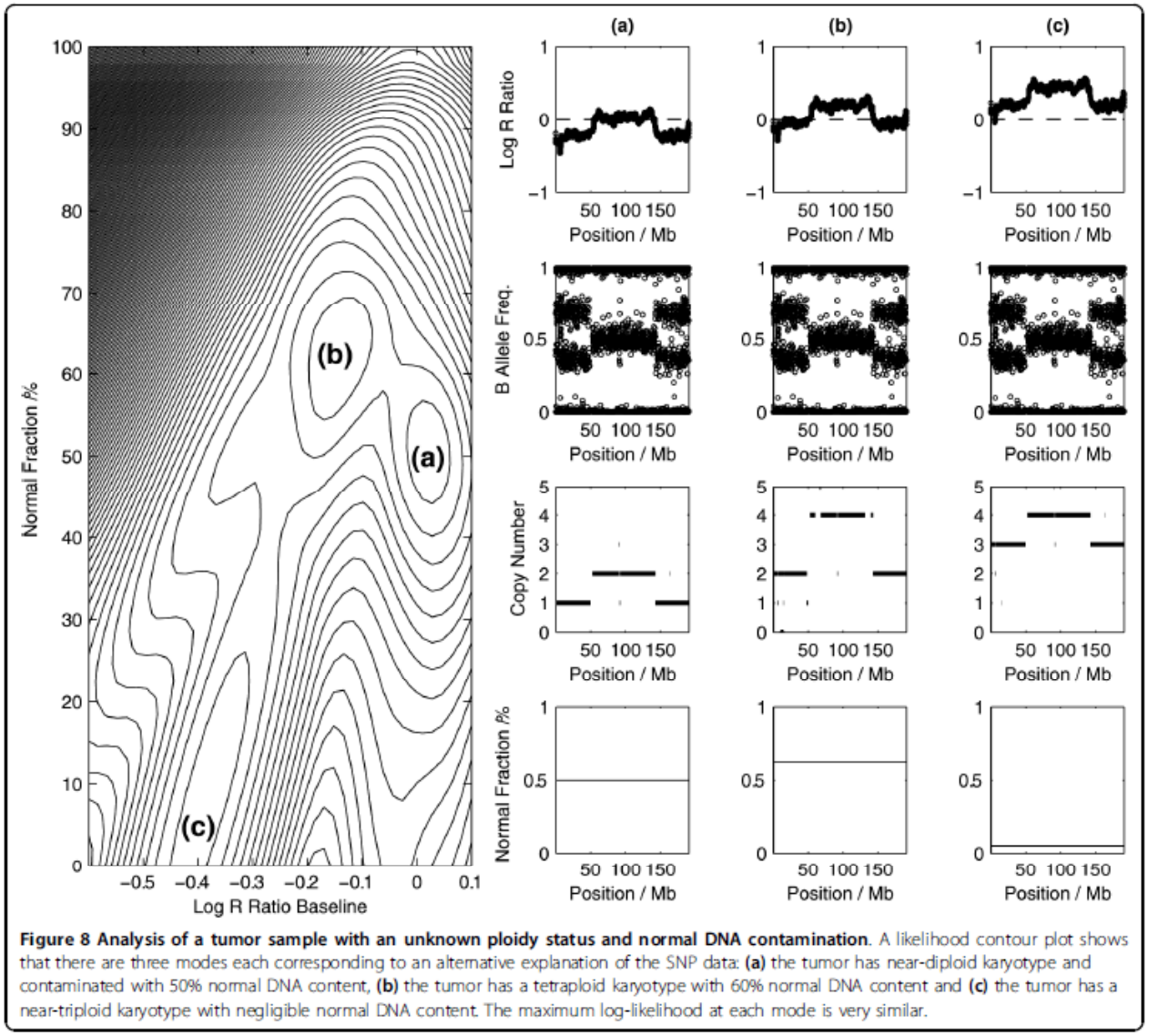


Figure 8 Analysis of a tumor sample with an unknown ploidy status and normal DNA contamination. A likelihood contour plot shows that there are three modes each corresponding to an alternative explanation of the SNP data: **(a)** the tumor has near-diploid karyotype and contaminated with 50% normal DNA content, **(b)** the tumor has a tetraploid karyotype with 60% normal DNA content and **(c)** the tumor has a near-triploid karyotype with negligible normal DNA content. The maximum log-likelihood at each mode is very similar.

How can one calculate %stromal contamination from dChip MCP, or compensate for %stromal contamination in dChip MCP scores, and correct MCP/LOH from contamination level particularly in paired tumors

- QiYuan's model, Bayesian approach
- ASCAT - PNAS paper model
- EM on dChip MCP or pre-MCP scores
- How does dChip work on so many platforms without "re-training" – it doesn't use training for any of its parameters

How can one calculate %stromal contamination from dChip MCP, or compensate for %stromal contamination in dChip MCP scores, and correct MCP/LOH from contamination level particularly in paired tumors

- 1) Build a model based on the dChip MCP/HMM model that incorporates a contamination model
- 2) Use BAF and Log R Ratio and train HMM states from those inputs ("dChip HMM on BAF")
- 3) Use an HMM to estimate contamination from copy numbers, BAF, LRR, instead of expectation-conditional maximization
- 4) Simple solution, variance of estimated MCP trained on tumor/stromal mixtures (see next slide)

HCC38M9 to HCC38M6 are mixture samples with tumor content of 90%, 80%, 70% and 60% respectively.

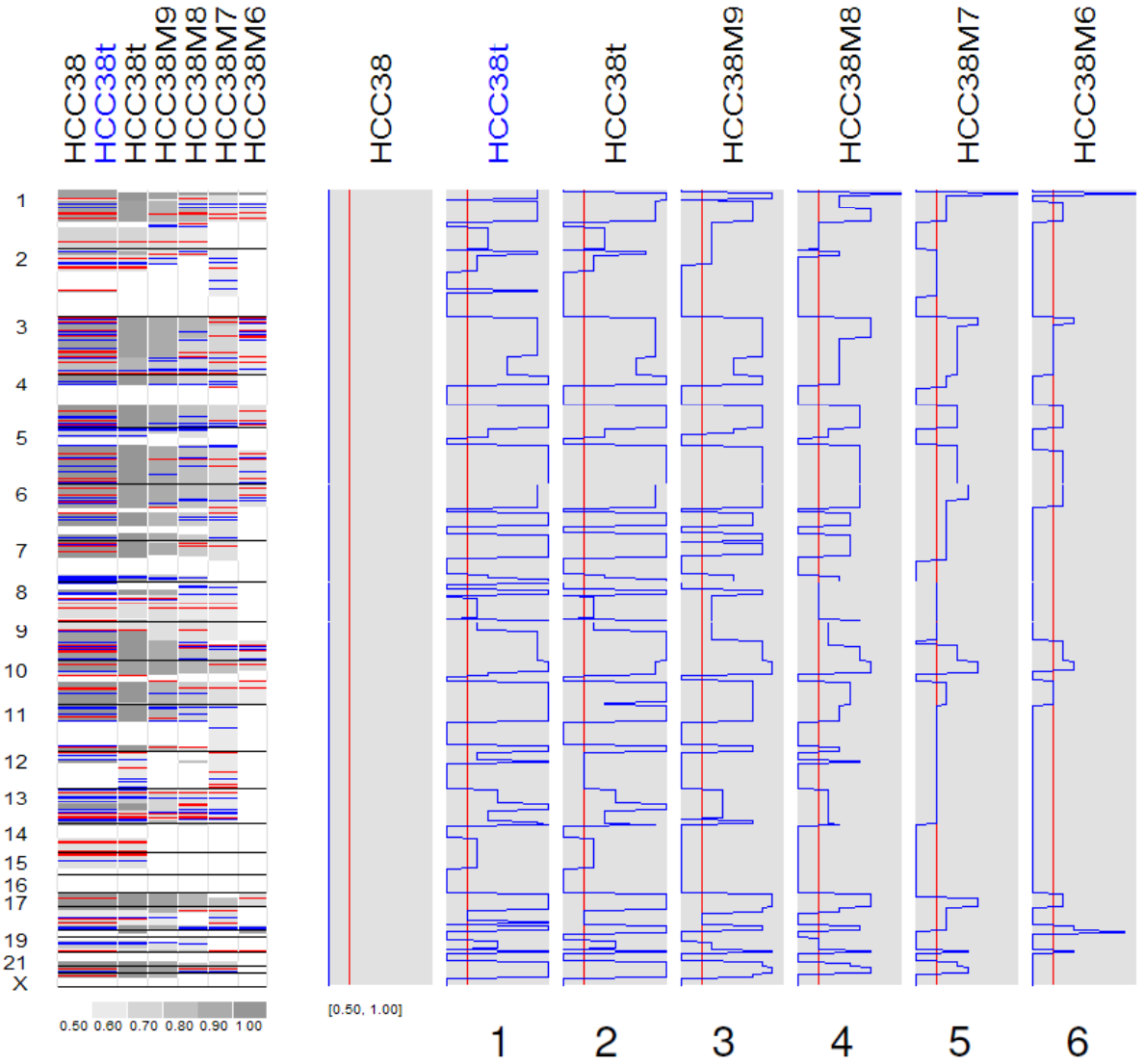


Figure 8
The genome-wide view of inferred MCP in the mixture samples. The red vertical lines in the gray boxes represent a MCP threshold of 0.6. See the legend of Figure 2B and 3D for color schemes.

Allele-Specific Copy number Analysis of Tumors ASCAT

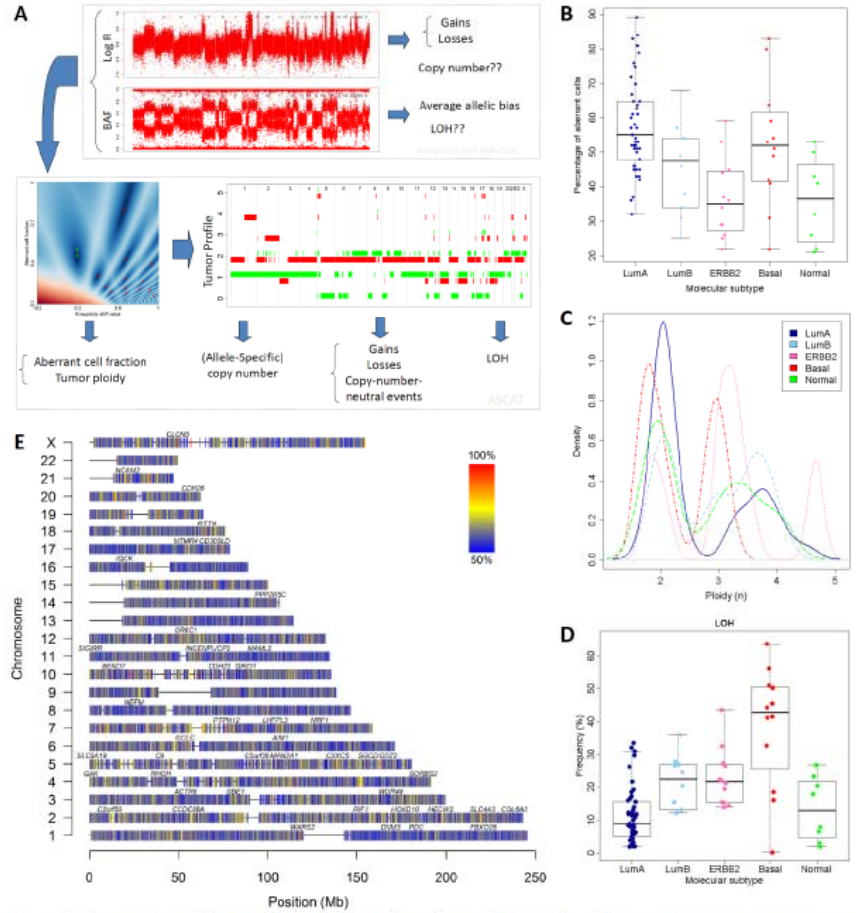


Fig. 1: (**A**) Tumor Profiles and the ASCAT algorithm. The results of an array-CGH or SNP-array experiment allow derivation of gains, losses and allelic bias. However, in cancer samples actual copy-numbers and LOH are difficult to determine, as tumors are often aneuploid and contain an unknown amount of non-aberrant cells. In this study, we developed ASCAT, a method to calculate accurate genome-wide allele-specific copy number profiles for tumor samples, taking into account tumor ploidy and non-aberrant cell admixture. An optimal ploidy and tumor percentage is determined and a Tumor Profile is calculated (red and green lines show the allele-specific copy number on the Y-axis vs. the genomic location on the X-axis; for illustrative purposes only, both lines are slightly shifted such that they do not overlap). These Tumor Profiles allow accurate derivation of gains, losses, copy-number-neutral events and LOH. (**B**) Percentage of aberrant tumor cells across the 5 molecular breast cancer subtypes. (**C**) Distribution of ploidy across the 5 subtypes. (**D**) Frequency of LOH per case, stratified by molecular breast cancer subtypes. (**E**) Genome-wide map of allelic skewness. Here, the frequency of the most frequently gained/lost allele is shown. Alleles without allelic skewness should have a frequency of 50 % (blue), while alleles that are completely skewed, have a frequency of 100 % (red). Gene symbols shown contain at least one SNP with a most frequently gained/lost allele frequency of 95 % or more.

$$\widehat{n}_{A,s}^{ASCAT} = \text{round}\left(\frac{\rho - 1 + 2^{\frac{r_s}{\gamma}}(1-b_s)(2(1-\rho)+\rho\psi_t)}{\rho}\right) \quad \textbf{[S10]}$$

$$\widehat{n}_{B,s}^{ASCAT} = \text{round}\left(\frac{\rho - 1 + 2^{\frac{r_s}{\gamma}}b_s(2(1-\rho)+\rho\psi_t)}{\rho}\right) \quad \textbf{[S11]}$$

where the *round()* function rounds to the nearest nonnegative integer. On the basis of these estimates $\widehat{n}_{A,s}^{\text{ASCAT}}$ and $\widehat{n}_{B,s}^{\text{ASCAT}}$, a theoretical Log R and BAF value ($\widehat{r}_s^{\text{ASCAT}}$ and $\widehat{b}_s^{\text{ASCAT}}$, respectively) is calculated:

$$\widehat{r}_s^{ASCAT} = \gamma\log_2\left(\frac{2(1-\rho)+\rho\left(\widehat{n}_{A,s}^{ASCAT}+\widehat{n}_{B,s}^{ASCAT}\right)}{2(1-\rho)+\rho\psi_t}\right) \quad \textbf{[S12]}$$

$$\widehat{b}_s^{ASCAT} = \frac{1-\rho+\rho\widehat{n}_{B,s}^{ASCAT}}{2-2\rho+\rho\left(\widehat{n}_{A,s}^{ASCAT}+\widehat{n}_{B,s}^{ASCAT}\right)}. \quad \textbf{[S13]}$$

Finally, both for Log R and BAF, an aberration reliability score ($l_{r,s}$ and $l_{b,s}$, respectively) is calculated as:

$$l_{r,s} = 1 - abs\left(\widehat{r}_s^{ASCAT} - r_s\right)/abs(r_s) \quad \textbf{[S14]}$$

$$l_{b,s} = 1 - abs\left(\widehat{b}_s^{ASCAT} - b_s\right)/abs(b_s - 0.5). \quad \textbf{[S15]}$$

In case of a copy number aberration without allelic imbalance [$abs(r_s) > 0.15$ and $b_s = 0.5$], the final aberration reliability score (percentage) l_s is given as $l_s = 100l_{r,s}$. In case of an allelic imbalance but no copy number aberration [$abs(r_s) \leq 0.15$ and $b_s \neq 0.5$; note that r_s and b_s are values obtained after ASPCF segmentation, which includes a check for one band with $b = 0.5$ vs. two bands symmetric around 0.5], $l_s = 100l_{b,s}$. In case of both a copy number aberration and an allelic imbalance [$abs(r_s) > 0.15$ and $b_s \neq 0.5$], $l_s = 50l_{r,s} + 50l_{b,s}$.

Hence, this aberration reliability score calculates for each aberration how well the ASCAT-predicted integer copy numbers match the data, compared with the hypothesis of no aberration. An aberration reliability score of 100% means ASCAT copy numbers perfectly explain the Log R and BAF data, whereas an aberration reliability score of 0 means the data are explained equally well by the ASCAT copy numbers as by the alternative hypothesis of no aberration.

ASCAT – from .no

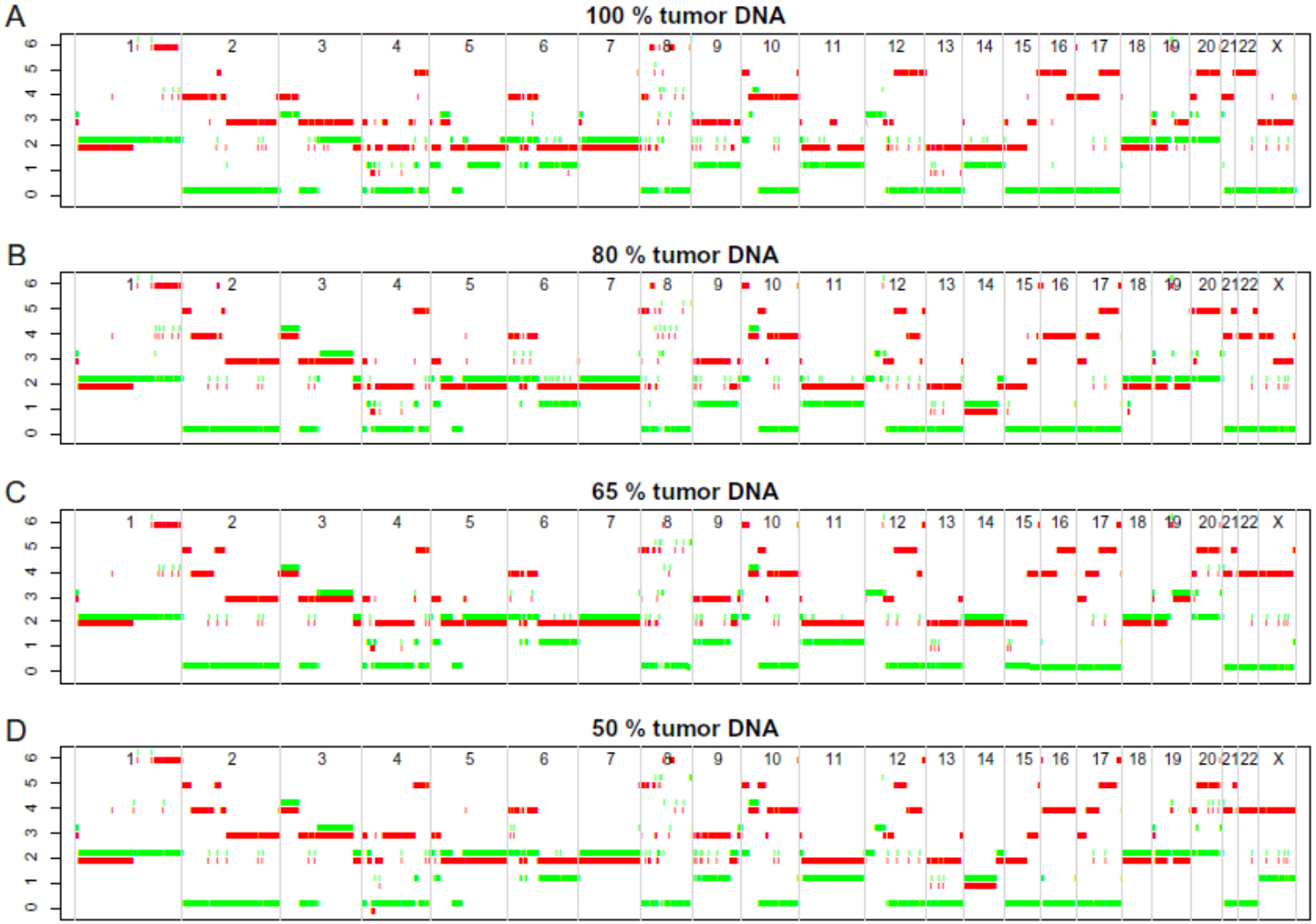


Fig. S4. Validation of ASCAT through a dilution series of a breast carcinoma. ASCAT profiles are shown for a dilution series of a highly aberrant breast carcinoma with ploidy 4.6n. Because the DNA mixes were produced by a total DNA weight ratio (i.e., not cell ratio), the annotated mixes correspond to (*A*) 100%, (*B*) 63%, (*C*) 45%, and (*D*) 30% aberrant tumor cells, assuming that the ploidy is close to 4.6n and the original tumor sample contained no nonaberrant cells. According to ASCAT, the samples contain (*A*) 83%, (*B*) 51%, (*C*) 46%, and (*D*) 32% aberrant tumor cells. Of all heterozygous probes, 64.8% (80% dilution), 60.3% (65% dilution), and 59.9% (50% dilution) showed exactly the same copy number for both alleles as the undiluted sample. For 95.3% (80% dilution), 93.9% (65% dilution), and 92.8% (50% dilution) of the heterozygous probes, the resulting copy numbers differ only slightly or not at all (a maximum copy number difference of 1 was allowed for each allele as well as for their sum).