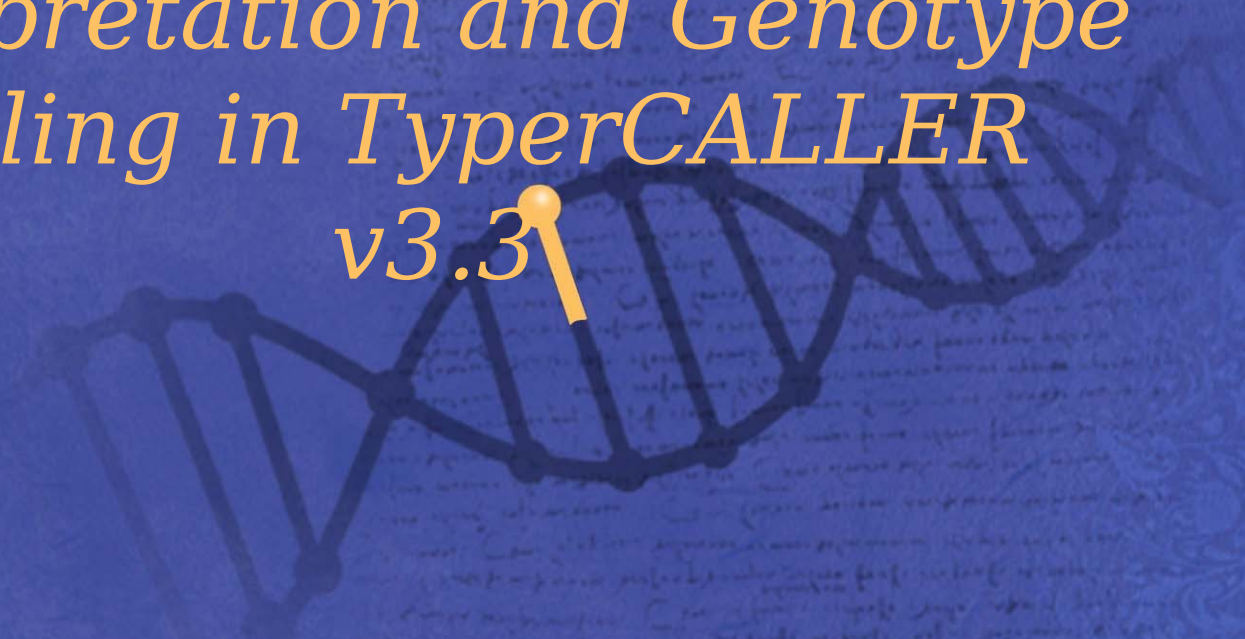


SEQUENOMTM

*Mass Spectra Signal
Interpretation and Genotype
Calling in TyperCALLER
v3.3*



Author: Dr. Guy Del
Mistro

1. Predicted linewidth
2. Concepts: signal; baseline; noise
3. Peak fitting details
4. Measurements from peak fitting
5. Assigning peak probability
6. Genotype calling

Appendix I: Probability functions

Appendix II: INI/Config parameters

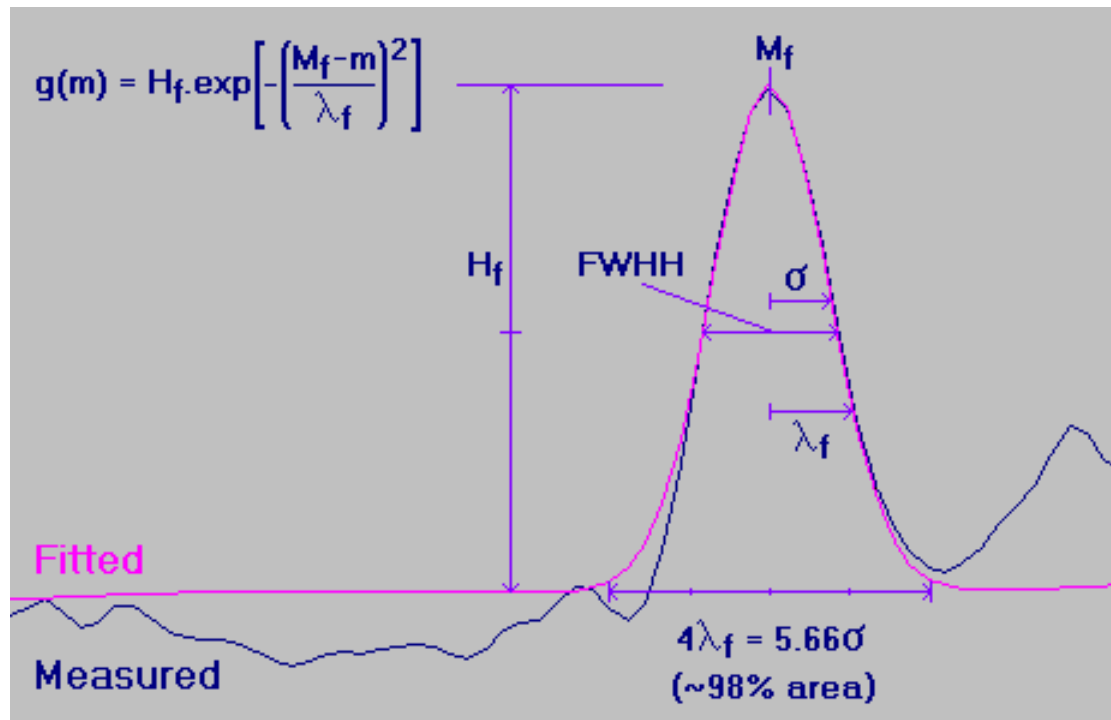
- In the absence of more specific information, the expected width of a mass signal (at a given mass) is fundamental to peak identification and quantification.
- Predicted linewidth (\approx HWHH) is specified for mass M by two INI parameters by the equation:

$$\lambda_e = L_A + L_B.M$$

- Initially (v3.1) λ_e was expected to be given by 0.1% of the signal mass (i.e. $L_A = 0$, $L_B = 0.001$). Later observations at lower and higher masses than usually considered in lower multiplexed spectra indicated that peak width did not vary so dramatically. For v3.3 (and SP3) more appropriate values for L_A and L_B were employed:

$$L_A = 2.5, L_B = 0.0005$$

- The (fitted) linewidth of a mass signal λ_f is defined by fitting a gaussian-shaped peak, $g(m)$, to that signal:



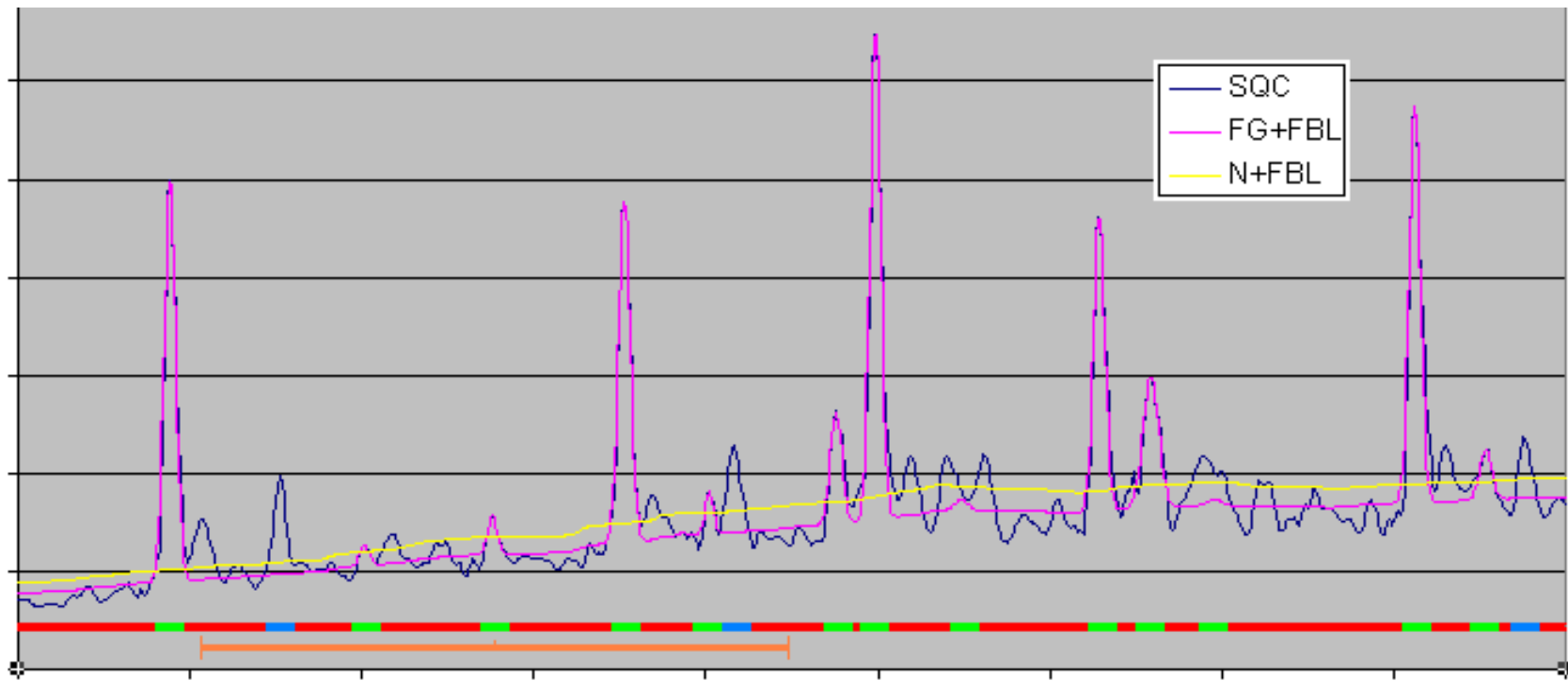
Comparison of Gaussian to the Normal Curve gives:

$$\lambda^2 = 2\sigma^2$$

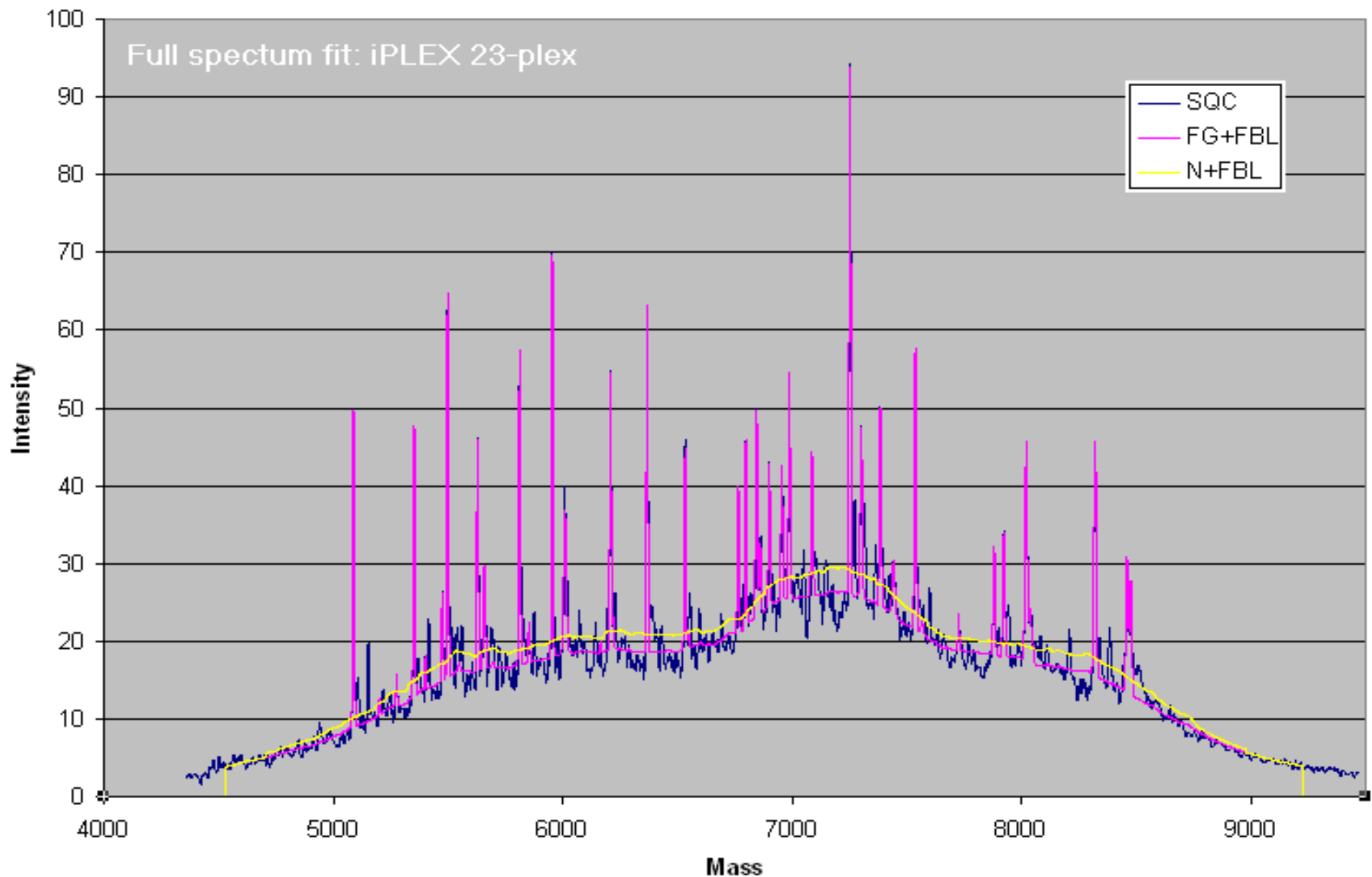
$$\begin{aligned} \text{FWHH} &= 2.3548\sigma \\ &= 1.6651\lambda \end{aligned}$$

- Reported signal resolution is defined from fitted linewidth:

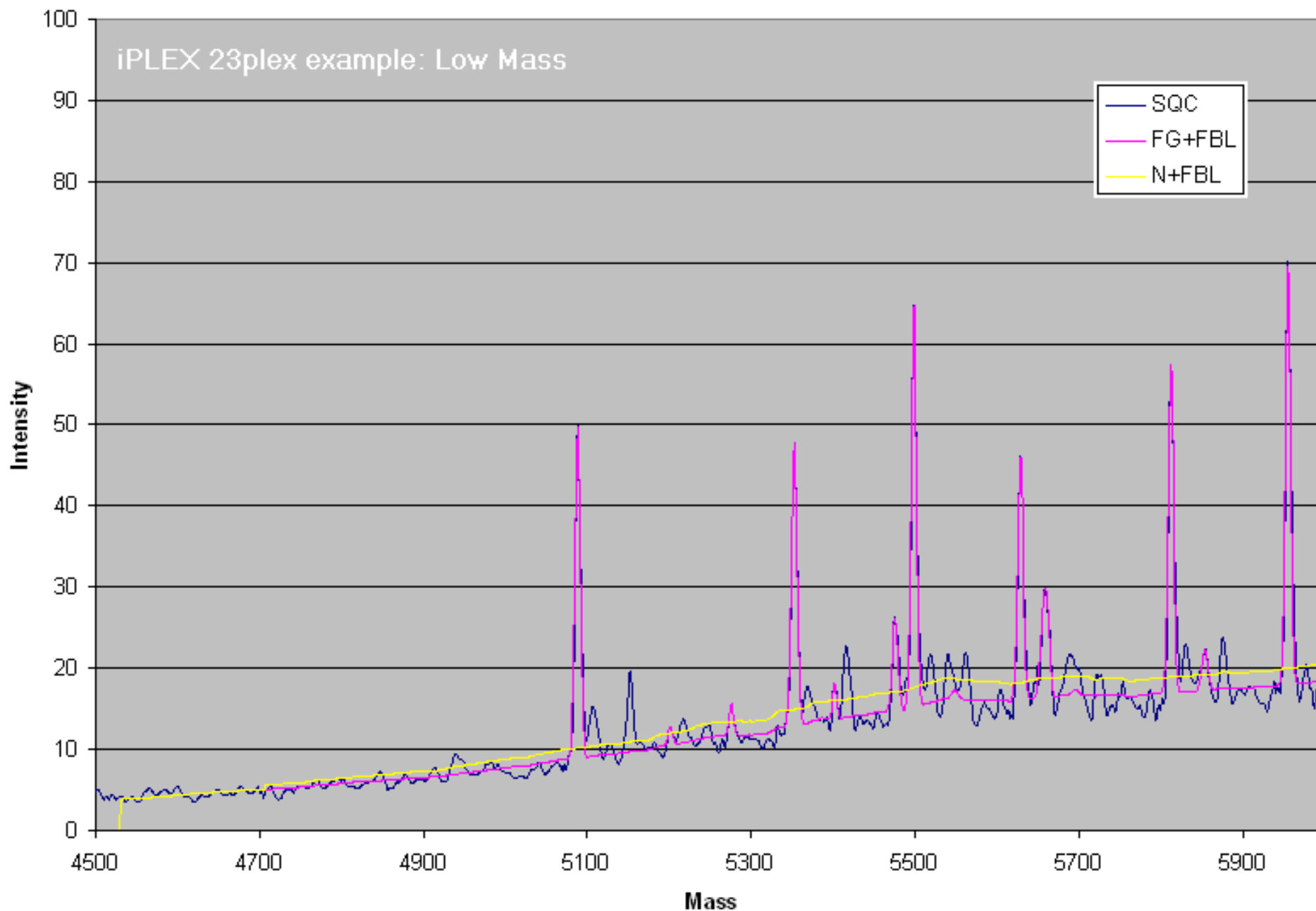
$$\text{Res}(M), M/\Delta M = M / \text{FWHH} = \underline{M / (1.6651 \cdot \lambda_f(M))}$$

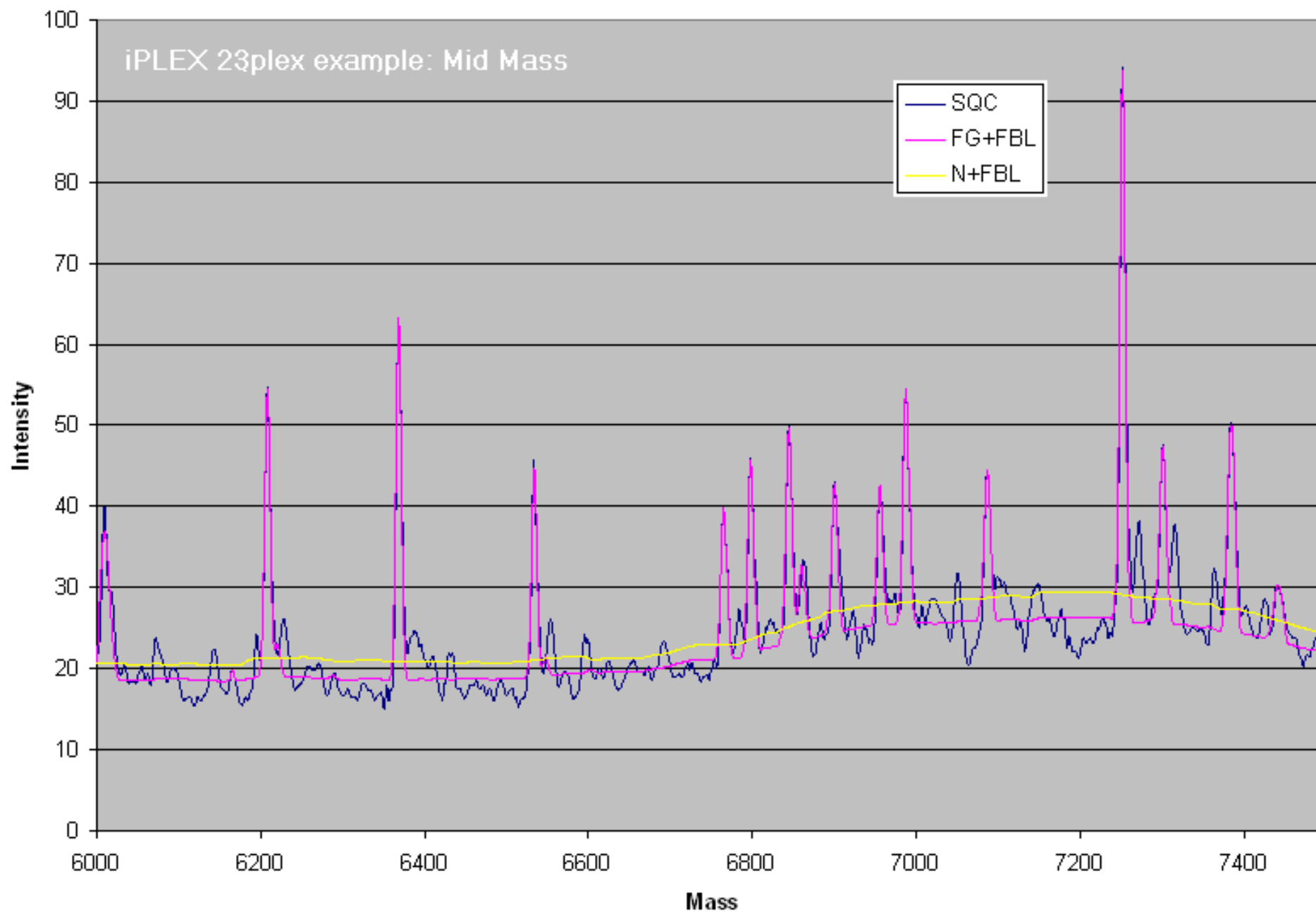


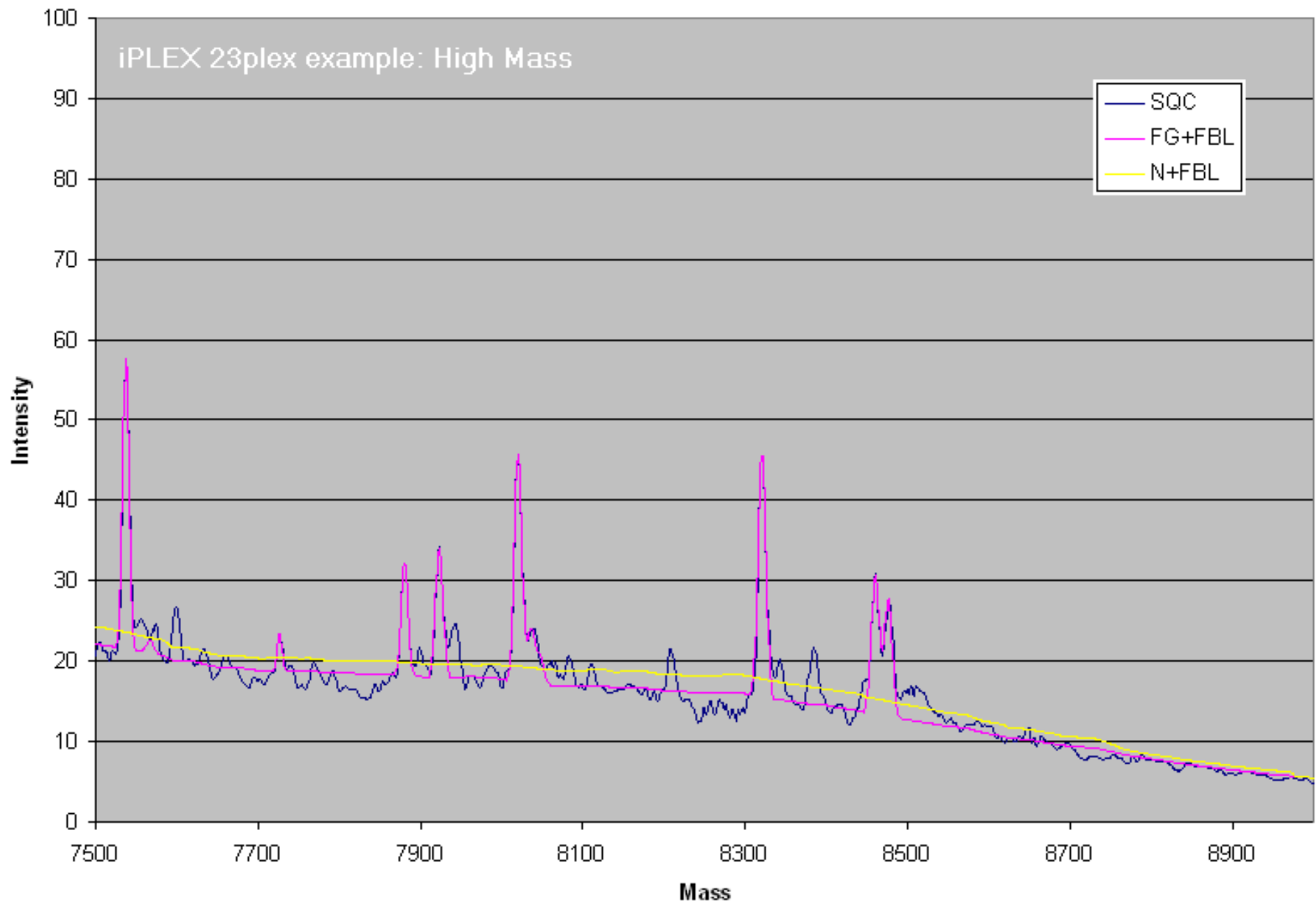
- Regions where signals are expected are masked out using $4\lambda_e$ (█)
- Baseline fitted using window of $80\lambda_m$ (█), where λ_m is the smallest λ_e (constant in point space, 400-540Da): *window width is important*
- Noise calculated as running RMS of (signal - baseline)
- Peaks discovered by masking points where height/noise (SNR) > 3 and noise > 1 . Repeated twice to avoid strongest contaminant peaks. (█)



Note 1st baseline point is N points in from ends; 1st noise point 2N in from ends



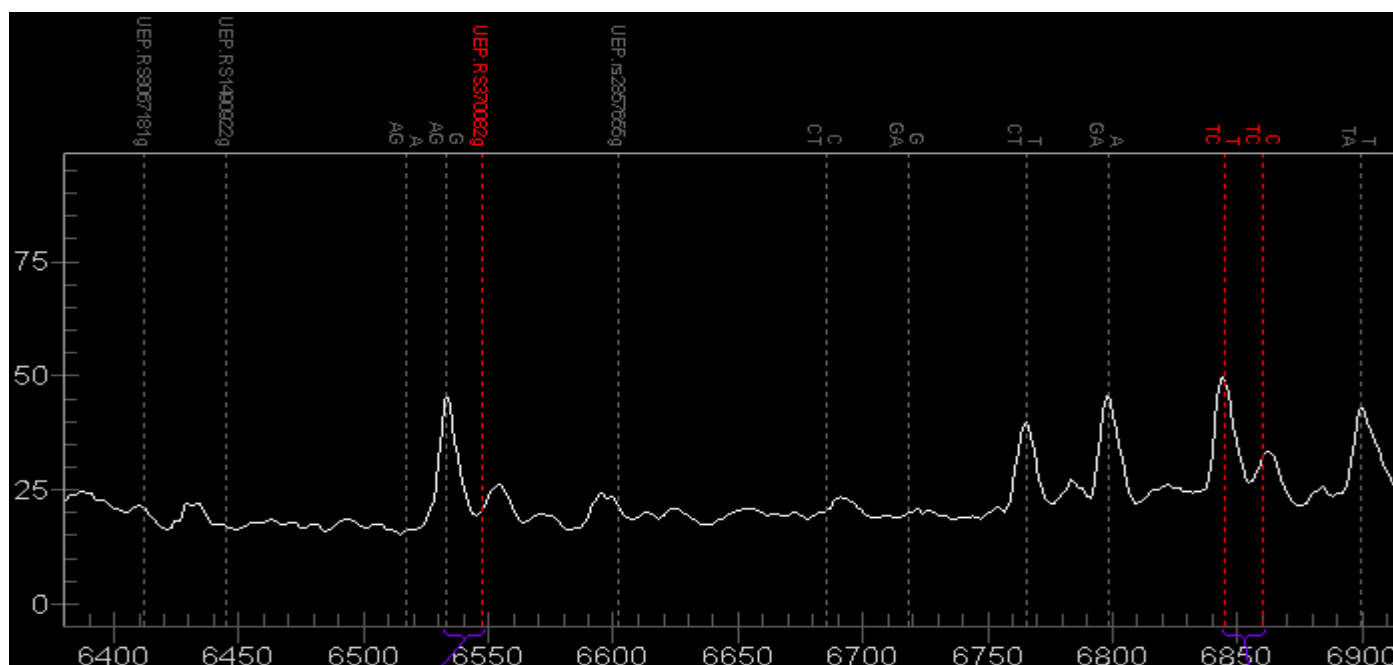




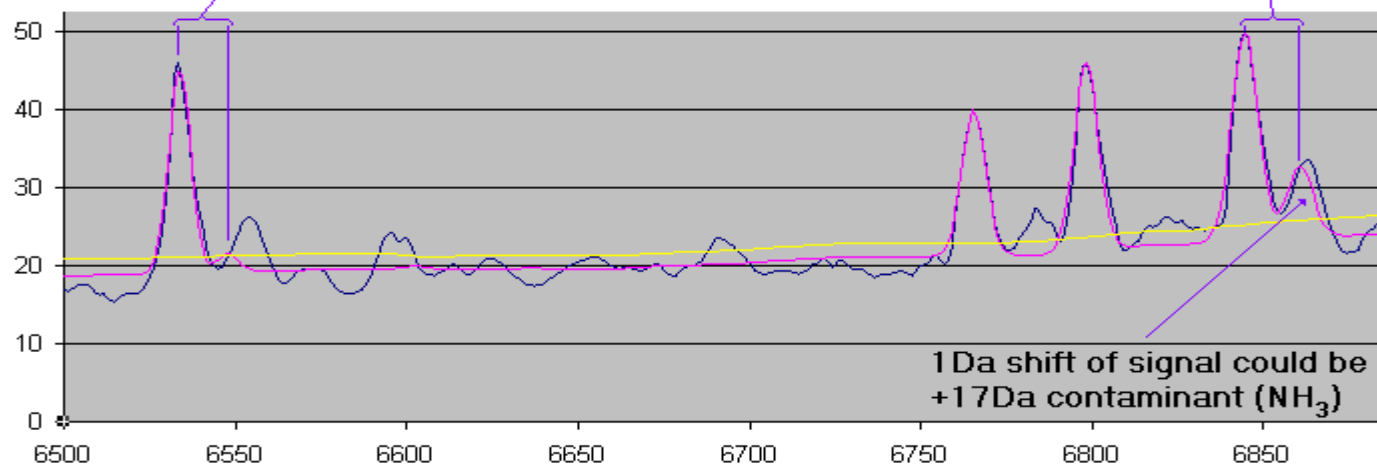
- One or more signals are fit to gaussian curves on the fitted baseline using the Levenberg-Marquardt minimization method.
- Multiple signals are fit as a group with a single phase if their expected masses (m_e 's) are within $4\lambda_e$ of each other.
- Each expected peak is fit over $4\lambda_e$ about the center but with 5x or 2.5x significance (depending on *peak type*) for the top λ_e (in data points).
- Fitting fails for a peak, or group of peaks, if:
 - $\lambda_f > 4\lambda_e$
 - $\lambda_f < \text{sampling resolution}$ (to catch data spikes)
 - phase (shift or *offset* of the peak center), $\delta_f > 3\lambda_e$
 - $H_f \leq 0$ (peak top is below the fitted baseline)

Example of multiple peak fit

SEQUENOM™

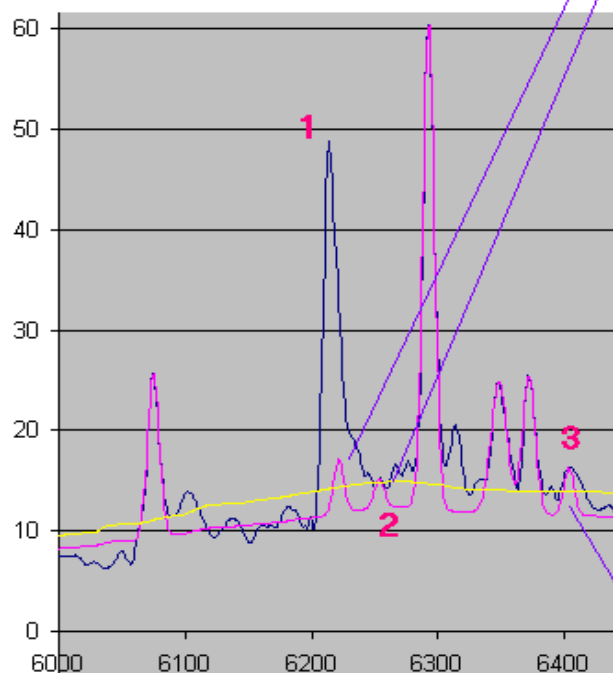
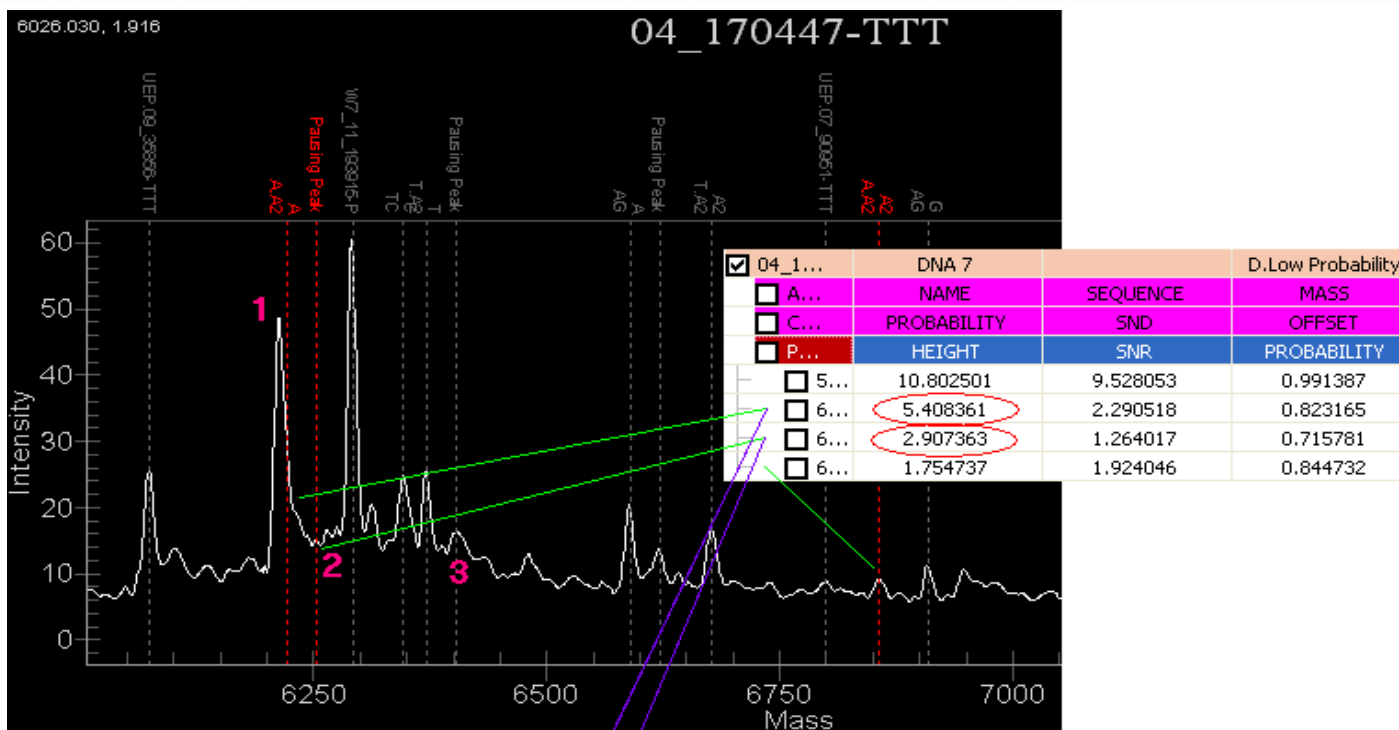


Peaks separated by less than 4 LW (23Da @ 6500Da) are fit to a set of gaussians at fixed mass spacing.



- Fitted linewidth is larger than a tolerance governed by an INI parameter, $\lambda_f > (1 + L_C) \cdot \lambda_e$, OR if the fitting fails:
 - $H(m_e)$ is used with the linewidth at λ_e .
 - or all measurements set to 0 if $H(m_e) \leq 0$.
- Offset is greater than a tolerance governed by an INI parameter, $\delta_f > L_C \cdot \lambda_e$:
 - Shape is further analyzed. A fitted contamination peak is subtracted to give height at m_e . (identifies shoulder peaks)
- Expected signals that are $< \lambda_e$ apart may be fit to a single peak as *any* linear combination of curves.
 - Heights and probabilities are adjusted with respect to their $H(m_e)$ values and peak type. For example, 2 primers at same mass fit to one peak would get $\frac{1}{2}$ of the height and ~ 0.5 prob.

Examples of obscured peak fitting



1 Large contaminant peak obscures analyte peak. Subtracted to give height at analyte mass (λ_e assumed).

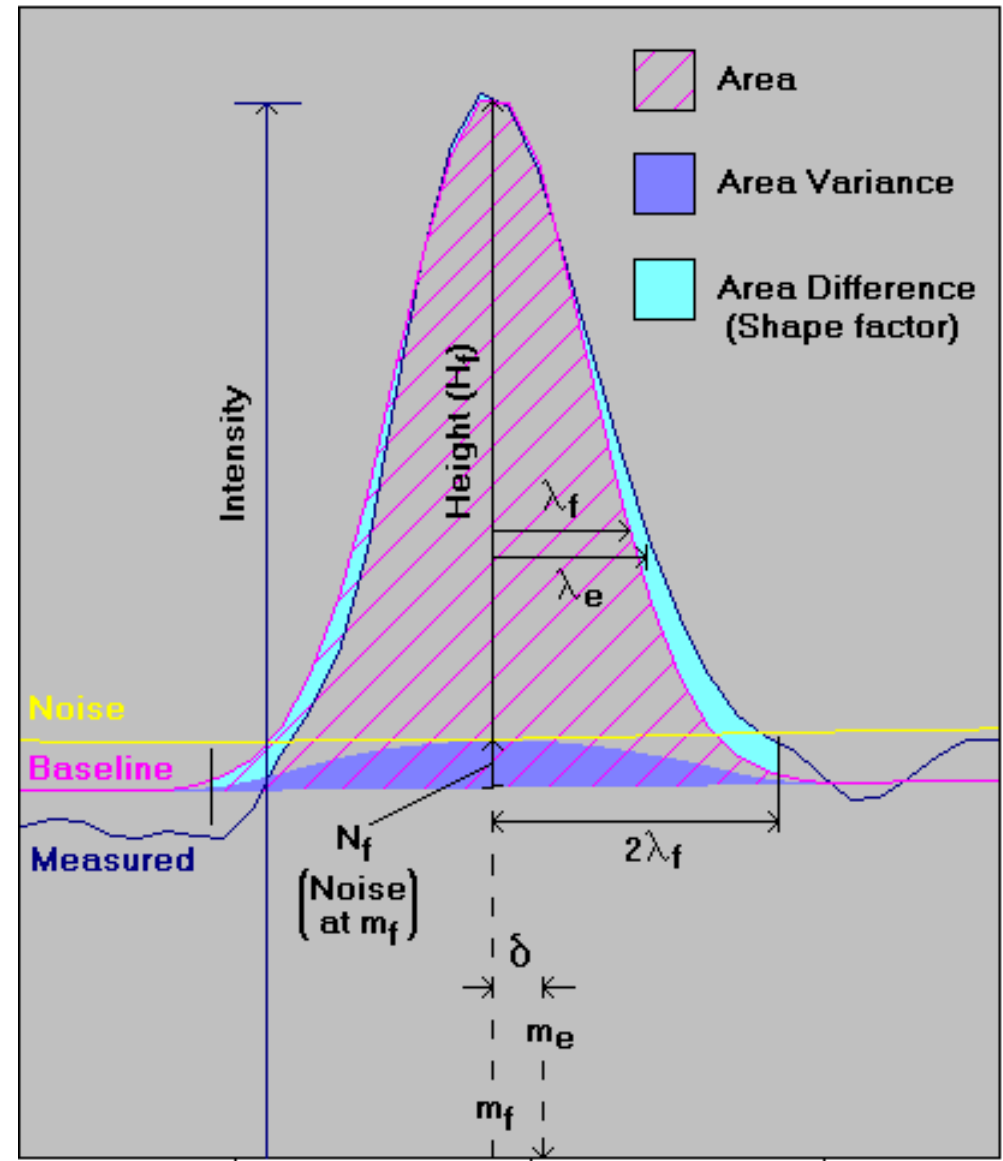
2 Top of peak enough to fit gaussian. Note height and is relative to fitted baseline.

3 Initial fit gave $\lambda_f > (1+C)\lambda_e$. Fitted height above the baseline is retained but linewidth is set to λ_e .

09_35856-TTT	DNA 7	T.A2	C.Aggressive
ASSAYINFO	NAME	SEQUENCE	MASS
CALLINFO	PROBABILITY	SND	OFFSET
PEAKINFO	HEIGHT	SNR	PROBABILITY
6074...	16.451437	9.995617	0.996635
6371...	13.807986	6.472106	0.989827
6403...	4.752717	2.140399	0.882075

4. Measurements from peak fitting

- Height above baseline, $H = H_f$
- Fitted linewidth, λ_f
- Offset: difference between fitted peak center and expected mass,
 $\delta = m_f - m_e$
- Area (A): integrated height above baseline under fitted peak over $4\lambda_f$
- Area Variance: uncertainty of peak area due to baseline fitting,
 $V = A \cdot H / N(m_f)$
- Fit area difference (Δ): integrated height difference between fit peak and measured signal over $4\lambda_f$
- SNR: peak height / baseline noise,
 $S = H / N(m_f)$



5. Assigning peak probability

- Overall peak probability is product of five components,

$$P = P_S \cdot P_{\Delta} \cdot P_{\delta} \cdot P_{\lambda} \cdot P_R$$

- P_S : (probability based on SNR)

- Indicates how distinct the peak is from local baseline variation (chemical noise)

- Main contributor for well fitted peaks

- P_{Δ} : (probability based on peak shape)

- Indicates how well a gaussian fits the measured peak shape

- May be significant contributor for poorly fitted peaks

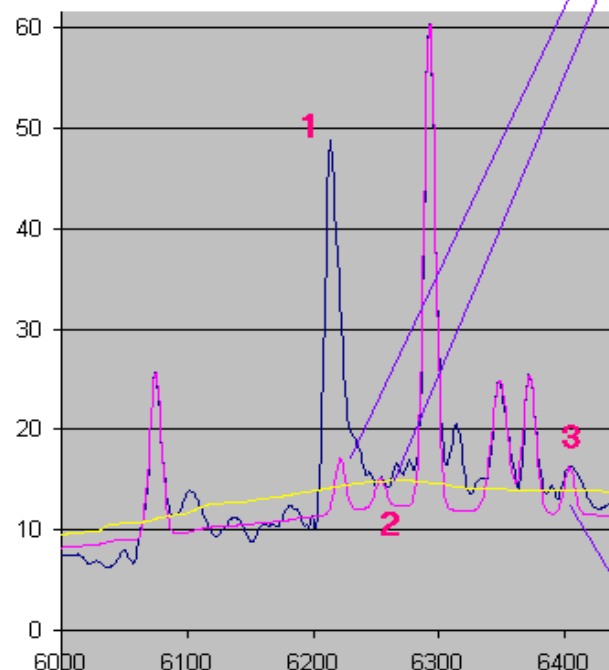
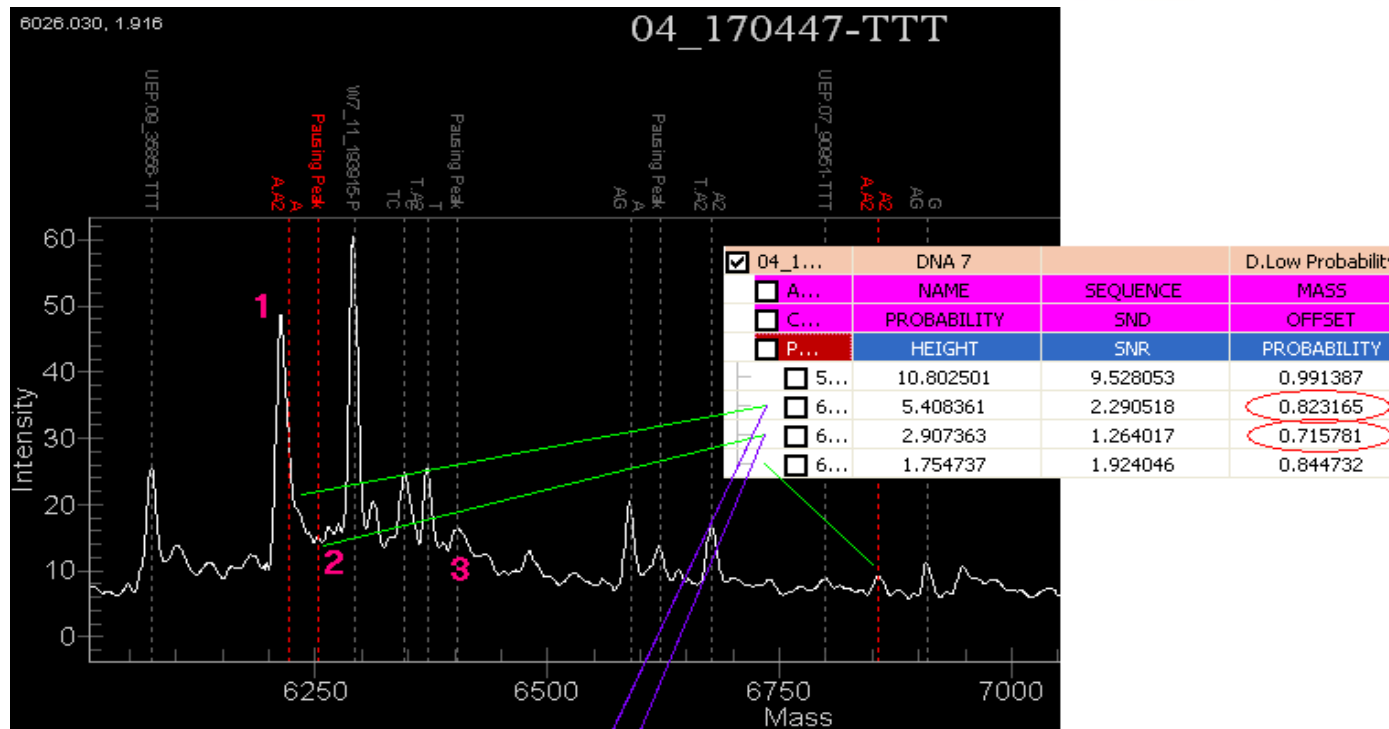
- P_{δ} : (probability based on peak offset)

- Indicates how close the peak was fitted to an expected mass signal

- Minor contributor: 1Da offset could be significant or just due to inaccuracy in the fitting or recalibration procedure

- P_{λ} : (probability based on peak width)

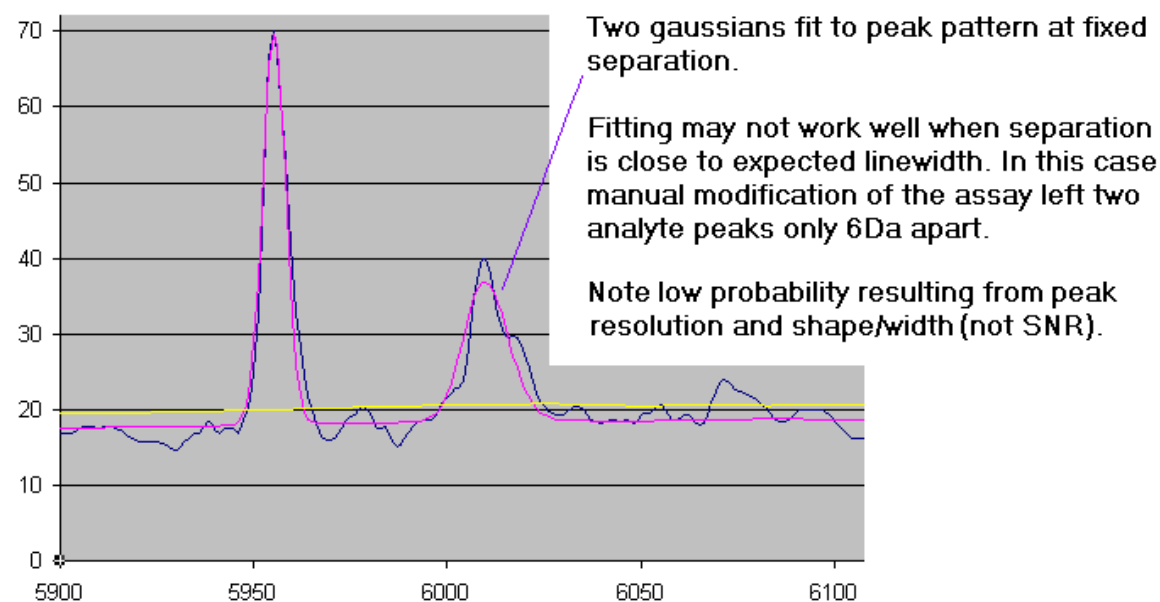
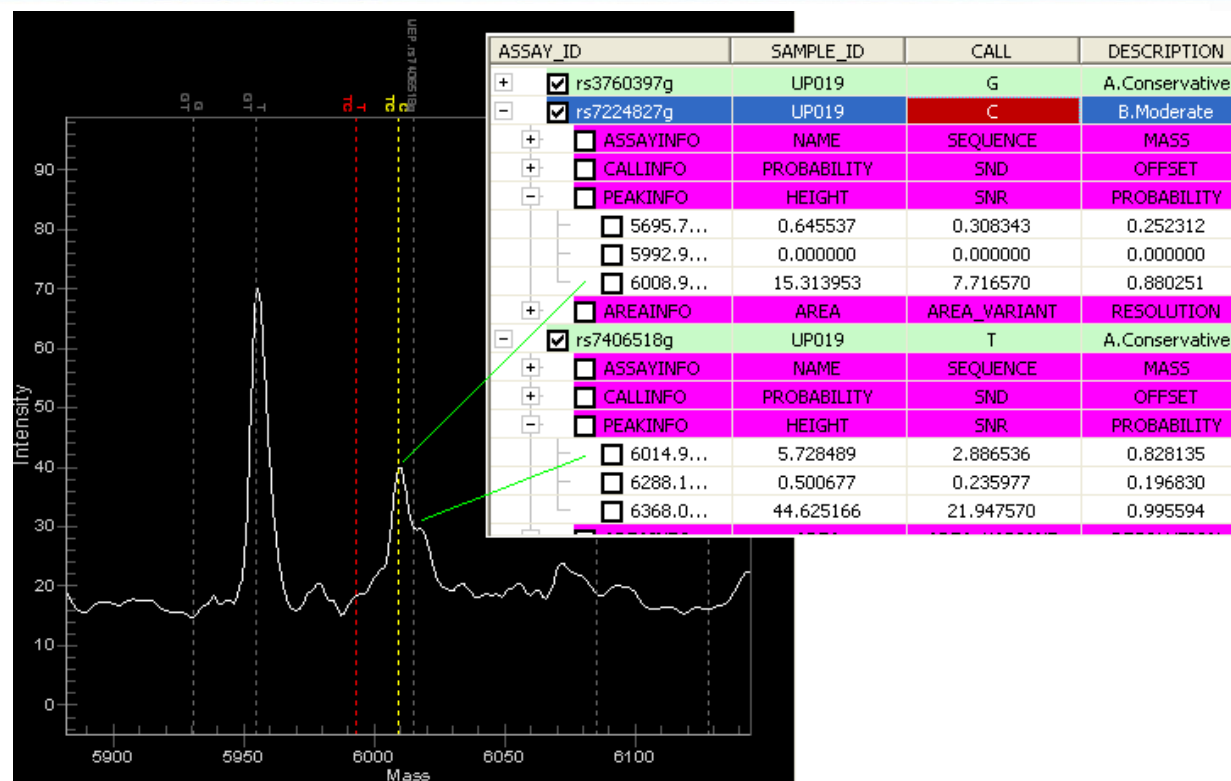
- Meant to trap broad peaks that have gaussian shape (mostly redundant)



- 1 The resulting obscured peak has a low probability partly due to low SNR but mainly because of area difference, P_A .
- 2 Peak has very low probability because it is largely below the noise and has a very low SNR.
- 3 Peak has moderate probability due to moderately low SNR and shape (area difference).

09_35856-TTT	DNA 7	T.A2	C.Aggressive
<input type="checkbox"/> ASSAYINFO	NAME	SEQUENCE	MASS
<input type="checkbox"/> CALLINFO	PROBABILITY	SND	OFFSET
<input type="checkbox"/> PEAKINFO	HEIGHT	SNR	PROBABILITY
<input type="checkbox"/> 6074...	16.451437	9.995617	0.996635
<input type="checkbox"/> 6371...	13.807986	6.472106	0.989827
<input type="checkbox"/> 6403...	4.752717	2.140399	0.882075

- P_R : (probability based on reference peak resolution)
- Only significant for masses *defined* to be close ($< 2\lambda_e$) to other assay masses.
- Takes into account *peak types*:
 - *Strong*: analyte, primer, specified contaminant
 - *Weak*: pausing, adduct, by-product
- An analyte mass defined too close to another assay mass (typically 4-7Da depending on the difference between peak types and masses) will result in a **Bad Assay** call.



- Genotype call primarily depends on the number of analyte peaks identified by their probabilities (P):
 - If all peaks have P's < C_{NOP} (no-peak cutoff), a **No Alleles** call is made.
 - If only one peak has $P > C_{\text{AGG}}$ (aggressive cutoff), a homozygous call may be made.
 - P_s is adjusted in expectation of a single analyte signal
 - Otherwise a call will depend on the relative skew of analyte signals
 - *Skew* is defined as the ratio of *signal intensities* for the smallest analyte to the largest analyte
 - Skew probability function adjusts probability of the minor allele using an INI parameter for the cutoff such that:
$$P_{\text{SKW}}(F_{\text{SKW}}) = C_{\text{AGG}} \Rightarrow \text{homozygous call made}$$

- Genotype *Call Score* is a product of three probability functions:

$$GCS = P_{MA} \cdot P_{SKW} \cdot P_{YLD}$$

- P_{MA} is the probability of the minor allele peak
- P_{SKW} is the skew probability: there are actually 4 such measures depending on whether forward or reverse skew is identified and whether the minor allele is ignored:
 - $P_{SKW} = 1$ if no second allele peak is identified
 - if $P_{SKW} < C_{AGG}$, P_{HSKW} is employed (new for Typer3.3)
- P_{YLD} is a probability factor based on the allele peak yield. An INI parameter specifies a minimum, such that:

$$P_{YLD}(F_{YLD}) = C_{AGG} \Rightarrow \text{Low Probability call made}$$

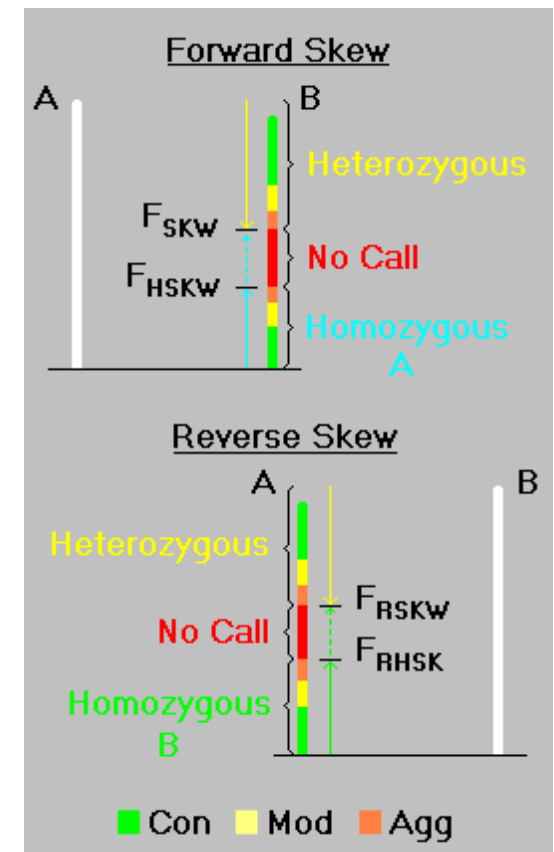
- GCS governs the confidence assigned a the genotype call:
 - $GCS > C_{CON} \Rightarrow \text{Conservative}$, $GCS > C_{MOD} \Rightarrow \text{Moderate}$
 - $GCS > C_{AGG} \Rightarrow \text{Aggressive}$, $GCS < C_{AGG} \Rightarrow \text{Low Probability}$

- Single Intensity for a given peak is represented as the product of the fitted height x fitted mass for that peak.
- hME allele peaks have a large mass separation
 - mass factor adjusts for width/height variation (minor)
 - SNR had issues due to different noise environments
- Skew = ratio of signal intensities
 - $= I_B / I_A, I_A \geq I_B$ (forward skew)
 - $= I_A / I_B, I_B > I_A$ (reverse skew)
- Heterozygous call IF

$$P_B \cdot P_{SKW}(I_B/I_A) > C_{AGG} \quad (\text{forward})$$
- ELSE Homozygous call if

$$P_B \cdot P_{HSKW}(I_B/I_A) > C_{AGG} \quad (\text{forward})$$
- Yield includes probe primer signal intensity:

$$YLD = (I_A + I_B) / (I_A + I_B + I_P)$$



SNR: $P_S = 1 - \exp(\alpha_S \cdot SNR), \quad \alpha_S = \ln(1 - C_{AGG}) / F_{SNR}$

Shape: $P_\Delta = \exp(-F_{SHP} \cdot \Delta / \lambda_f \cdot H_f)$

Shift: $P_\delta = \exp(\alpha_\delta \cdot (\delta / \lambda_e)^4), \quad \alpha_\delta = \ln(C_{AGG}) / (L_C)^4$

Width: $P_\lambda = 1 - \exp(\alpha_\lambda \cdot (\lambda_e - \lambda_f)^2), \quad \alpha_\lambda = -0.01 F_{WTH}$

RPR: $P_R(j) = \prod_{i \neq j} \{ 1 - \exp(\alpha_R (m_i - m_j)^2) \cdot \tau_i / (\tau_i + \tau_j) \},$
 - m_i and τ_i are the mass and peak type weighting for
 the i^{th} expected assay peak (*strong*: 1, *weak*: 0.05),
 $\alpha_R = \ln(0.02) / (2.5 L_C \cdot \lambda_e(j))^2$

Skew: $P_{SKW} = (1 - \exp(\alpha_{SKW} \cdot S)) / W_S, \quad \alpha_{SKW} = \ln(1 - C_{AGG}) / F_{SKW}$
 $W_S = 1 - \exp(\alpha_{SKW})(S - F_{SKW}) / (1 - F_{SKW}) \Rightarrow P_{SKW}(S=1) = 1$

$P_{HSKW} = \exp(\alpha_{HSKW} \cdot S), \quad \alpha_{HSKW} = \ln(C_{AGG}) / (F_{HSKW} \cdot F_{SKW})$

$S = I_B / I_A$: for reverse skew $S = I_A / I_B$; substitute F_{RSKW} and F_{RHKS}

Yield: $P_{YLD} = 1$ if $YLD > Y_G$, $Y_G = 0.7$ (70%), otherwise:

$P_{YLD} = \exp(\alpha_{YLD} \cdot YLD / (Y_G - YLD)), \quad \alpha_{YLD} = \ln(1 - C_{AGG}) \cdot (1 - F_{YLD}) / F_{YLD}$

L_A	PEAKFIT_LINEWIDTH_A	2.5	Minimum expected linewidth
L_B	PEAKFIT_LINEWIDTH_B	0.0005	Expected linewidth broadening with increasing mass
L_C	PEAKFIT_AGG_SHIFT_F	0.7	Offset tolerance relative to expected peak linewidth
F_{SNR}	PEAKFIT_AGG_SNR_T	1.5	SNR threshold for peak identification
F_{SHP}	PEAKFIT_GAU_SHAPE_F	0.2 (0.1)*	Penalty factor for peak shape vs. fitted gaussian
F_{WTH}	PEAKFIT_GAU_WIDTH_F	0.1 (0.05)	Penalty factor for fitted peak width vs. expected
C_{NOP}	GENOTYPE_NOP_CUTOFF	0.7	Initial probability cutoff for peak identification
C_{AGG}	GENOTYPE_AGG_CUTOFF	0.8	Call score cutoff for aggressive call status
C_{MOD}	GENOTYPE_MOD_CUTOFF	0.85	Call score cutoff for moderate call status
C_{CON}	GENOTYPE_CON_CUTOFF	0.93	Call score cutoff for conservative call status
F_{YLD}	GENOTYPE_AGG_YSKW_T	0.1 (0.25)	Extension yield threshold (for aggressive call)
F_{SKW}	GENOTYPE_AGG_SKEW_T	0.2 (0.5)	Forward skew threshold for (agg) HET / No Call
F_{RSKW}	GENOTYPE_AGG_RSKW_T	0.3 (0.5)	Reverse skew threshold for (agg) HET / No Call
F_{HSKW}	GENOTYPE_AGG_HSKW_T	0.75 (0.6)	Forward skew threshold for No Call / HOM (agg)
F_{RHSK}	GENOTYPE_AGG_RHSK_T	0.75 (0.6)	Reverse skew threshold for No Call / HOM (agg)

* Values in parentheses are for iPLEX where they differ from those for hME

PEAKFIT_RECALMAXSHIFT 10.0 Maximum allowed spectrum recalibration shift (Da)

PEAKFIT_BASE_NOISE 0.15 Added to noise to prevent large SNR in smooth regions