Illumina Omniチップデータからの 体細胞ゲノムコピー数異常予測ソフトウエアの調査

谷村直樹[i] 知久季倫[i] 新井恵吏[ii] 金井弥栄[iii] 坂本裕美[iv] 吉田輝彦[v]

A research of somatic DNA copy number aberration prediction sotfware from Illumina Omni SNPs chip data

Naoki tanimura Suenori Chiku Eri Arai
Yae Kanai Hiromi Sakamoto Teruhiko Yoshida

Illumina の SNPs チップの log R ratio 及び B allele frequency データから，somatic なコピー数変異である copy number somatic abberations/alterations (CNA) を推定するソフトウエアの調査を行った．この結果 ASCAT[4] と GPHMM[2] の性能が高いと判断し，両ソフトウエアによるパイプラインを構築した．

(**キーワード**): **がん**, DNA, CNV, CNA, SNPs, Omni

1 はじめに

Germline の遺伝子型をチップを使って大量に測定する方法は，比較的安価で且つ高い信頼性があることが広く認められ，様々な疾患リスクを予測する商業的な使用を含め普及している．近年，この SNPs チップを遺伝子型タイピングの枠を越えて，array comparative genomic hybridization (aCGH) データとしてゲノムコピー数の推定に用いる研究が数多く報告され，更にがん細胞における somatic なコピー数変異である copy number somatic abberations/alterations (CNA) の推定にも利用されている．がん細胞における CNA の同定は，特定の遺伝子の loss of heterozygosity (LOH) や増幅が古くから研究されている様に，がんの driver mutation の探索における知見になると考えられている．更に DNA のメチル化や次世代シーケンサー (NGS) を使ったがんの single nucleotide variation (SNV) 探索において，がん臨床検体に混入している非がん部の割合を求められることがあり，NGS に比べて安価な SNPs チップによる測定が行われることがある．

そこで本報告書では，Illumina の SNPs チップによるデータから非がん部の混入を考慮した CNA 予測ソフトウエアの調査を行い，更にこれらのソフトウエアを実データに適用させて評価を行い，CNA 予測のためのパイプラインを構築することを目的とした．

[i]サイエンスソリューション部 バイオエンジニアリングチーム
[ii]国立がん研究センター 研究所 分子病理分野 主任研究員
[iii]国立がん研究センター 研究所 分子病理分野 分野長
[iv]国立がん研究センター 研究所 遺伝医学研究分野 ユニット長
[v]国立がん研究センター 研究所 遺伝医学研究分野 分野長

2 文献調査

2.1 Somatic 変異のコピー数推定

2.1.1 Sun *et al.* (2010)[6](GenoCNA)

genoCNA は，正常組織のコンタミネーションを考慮してがん細胞の copy number variations (CNVs) および copy number aberrations (CNAs) を分析するためのプログラムである genoCN の構成要素 (genoCNV, genoCNA) の 1 つである． genoCN は，Illumina の SNP アレイから出力される Log R ratio (LRR) と B allele frequency (BAF) を入力とし，がん細胞での各々の SNP のコピー数と遺伝子型の事後確率を出力する．

genoCNV と genoCNA は，CNV と CNA データが異なる遺伝子型クラスであること，及び CNA データでの正常組織のコンタミネーションの影響を考慮して異なる設計がなされている．ここで述べる genoCNA では同一患者から取得された正常組織の遺伝子型を入力に用いることが可能であり，分析結果の頑健性と精度を高めることが可能である．genoCNA では，表 1 に記載した 4 コピー数までをカバーする 9 state の hidden Markov model(HMM) を用いて予測を行う．

染色体上での並びに対応した $i(1 < i < L)$ 番目の SNP プローブについて，LRR，BAF，隠れ状態，及び正常組織から得られた遺伝子型をそれぞれ r_i，b_i，z_i 及び g_i とすると尤度は

$$p(r_1, \cdots, r_L, b_1, \cdots, b_L) = \sum_{z_1, \cdots, z_L} \left[p(z_1) \prod_{i=1}^{L} p(r_i|z_i) \prod_{i=1}^{L} p(b_i|z_i, g_i) \prod_{i=2}^{L} p(z_i|z_{i-1}) \right]$$

表 1 genoCNA で使われている染色体の 9 状態. 表示例:(A,AB) は，がん組織の遺伝子型 A と正常組織からの遺伝子型 AB のコンタミネーション.

State ID	Copy #	Description	genotype
1	2		AA, AB, BB
2	2	only homozygous	AA, (AA,AB), (BB,AB), BB
3	0	both alleles are deleted	Null
4	1	hemizygous	(A,AA), (A,AB), (B,AB), (B,BB)
5	3		(AAA,AA), (AAB,AB), (ABB,AB), (BBB,BB)
6	3	only homozygous, 1 allele is deleted first before the other is amplified	(AAA, AA), (AAA,AB), (BBB,AB), (BBB,BB)
7	4	simultaneous amplification of both alleles	(AAAA,AA), (AABB,AB), (BBBB,BB)
8	4	only homozygous, 1 allele is deleted first before the other is amplified	(AAAA,AA), (AAAA,AB), (BBBB,AB), (BBBB,BB)
9	4	amplification of one allele twice	(AAAA,AA), (AAAB,AB), (ABBB,AB), (BBBB,BB)

となる. ここで $p(z_i|z_{i-1})$ は隠れ状態間の遷移確率である. また $p(r_i|z_i)$ 及び $p(b_i|z_i, g_i)$ はそれぞれ LRR と BAF の出力確率 (emission probability) であり，隠れ状態が与えられたときに独立であると仮定されている. 更に正常組織のコピー数は 2 でありがん細胞のコピー数についての情報をもたらさないこと，正常組織の遺伝子型はがん細胞の遺伝子型に影響を与えることが仮定されている.

遷移確率

遷移確率は $i-1$ 番目のプローブと i 番目のプローブの距離を d_i として，

$$p(z_i = k|z_{i-1} = j) = \begin{cases} e^{-\lambda_j d_i} & \text{for } k = j \\ a_{jk}(1 - e^{-\lambda_j d_i}) & \text{for } k \neq j \end{cases}$$

と表されている. ここで $a_{jk}(k \neq j)$ は，j から j 以外の状態に遷移する場合の遷移確率であり，$\sum_{k \neq j} a_{jk} = 1$ である. λ_i は事前知識等によりユーザーが設定するパラメーターである. 隣り合う SNPs プローブ間での状態遷移は 1 回以内とする仮定を置いており，またプローブ間の距離が離れている場合には Markov 過程を再スタートさせることにしている.

LRR の出力確率

LRR(r) の出力確率は一様分布及び平均 μ，標準偏差 σ の正規分布 $\phi(r; \mu, \sigma)$ の混合により

$$p(r|z) = \pi_{r,z}\frac{1}{R_m} + (1 - \pi_{r,z})\phi(r; \mu_{r,z}, \sigma_{r,z})$$

と表現されている. ここで $\pi_{r,z}$ は隠れ状態 z での一様成分の混合比率，R_m は LRR の範囲を表す長さ，$\mu_{r,z}$ と $\sigma_{r,z}$ はそれぞれ隠れ状態 z における LRR シグナルの平均値と標準偏差である.

BAF の出力確率

BAF(b) の出力確率についても一様分布及びパラメーター $\theta = \{\mu$:平均,σ:標準偏差 $\}$ の正規分布 $\phi(b; \theta)$(累積分布 $\Phi(b; \theta)$) の混合により

$$\begin{aligned} p(b|z) =& \pi_{b,z} I(0 < b < 1) \\ &+ (1 - \pi_{b,z}) \sum_{h=1}^{H_z} \Big[w_{z,h} \phi(r; \theta_{z,h})^{I(0<b<1)} \\ &\times \Phi(r; \theta_{z,h})^{I(b=0)} \{1 - \Phi(r; \theta_{z,h})\}^{I(b=0)} \Big] \end{aligned}$$

と表されている. ここで $\pi_{b,z}$ は隠れ状態 z での一様成分の混合比率，$I(\cdot)$ は指示関数，H_z は隠れ状態 z での全成分 (遺伝子型) の数，w_h はその h 番目の成分の重み，$\theta_{z,h} = \{\mu_{b,z,h}, \sigma_{b,z,h}\}$ は h 番目の成分の平均と標準偏差である (遺伝子型は B allele の数の順に並べる).

w_h は，正常組織の遺伝子型についての情報がない場合には，B allele の集団中での頻度に基づいた 2 項分布等によって決定される. 正常組織の遺伝子型の情報が得られる場合には，タイピングエラー率も含めさらに精緻化された重みづけと出力確率の設定が行われる (Supplementary Materials B.1 に詳述されている).

プログラム入出力，モデルパラメータ推定

入力データは $\mathbf{X} = \{x_i, r_i, b_i, \lambda_j, g_i\}$ である．ここで x_i はプローブ位置であり，g_i はオプションである．

推定すべきパラメータは，

$$\Theta = \{\pi_{r,z}, \mu_{r,z}, \sigma_{r,z}, \pi_{b,z}, \mu_{b,z,h}, \sigma_{b,z,h}, a_{jk}\}$$

であり，Baum-Welch アルゴリズムによって推定を行っている (Supplementary Materials B.3–7 に詳述)．尚，幾つかのパラメーターについてはコピー数や遺伝子型の同一性により同一の値とし，また正規分布の平均値については状態設定に基づき定数としているものもある．

推定されたパラメータに基づいて計算される各々の SNP に対するコピー数及び遺伝子型の状態の事後確率が最終的な出力である．

プログラム実装

genoCN プログラムは R のパッケージとして実装されており，計算負荷の大きい個所については C 言語で記述されている．

2.1.2 Loo *et al.* (2010)[4] (ASCAT)

ASCAT (allele-specific copy number analysis of tumors) は，Illumina の SNPs チップから出力されるがん細胞と非がん細胞それぞれの log R ratio (LRR) と B allele frequency (BAF) から，がん細胞と非がん細胞の混合比率とがん細胞の平均 ploidy を推定し，A allele と B allele のコピー数を推定する．推定手順の概要は，

1. 非がん部のデータからヘテロの SNPs を抽出し，この SNPs を使ってがん部のデータをセグメント化する (allele-specific piecewise constant fitting; ASPCF).
2. セグメント結果からがん部と非がん部の混合比率とがん細胞の平均 ploidy を推定.
3. Allele 毎のコピー数を推定し，reliability score を計算.

である．

がん細胞のセグメント推定 (ASPCF) では，染色体上の位置 (x_i) 順に並んだ n 個の SNPs データから得られる LRR の集合を $\{(x_i, r_i), i = 1, \cdots, n\}$，BAF の集合を $\{(x_i, b_i), i = 1, \cdots, n\}$ と表したとき，

$$\begin{aligned} l &= \sum_{j=1}^{Q} \sum_{i \in I_j} \left[w\{r_i - ave(r_s)_{s \in I_j}\}^2 \right. \\ &\quad \left. + (1-w)\{b_i - ave(b_s)_{s \in I_j}\}^2 \right] + \lambda Q \end{aligned}$$

が最小になるように Q 個のセグメントの集合 I_j を決める．ここで重み w はデフォルトで 0.5，$ave(r_s)_{s \in I_j}$ はセグメント I_j 領域内での平均値，ペナルティ項の重み $\lambda > 0$ はデフォルト $\lambda = 50$ としている．また最小セグメント長として 6 SNPs を条件としている．

がん部と非がん部の比率を $\rho : (1-\rho)$ とし，がん部の主要な ploidy を Ψ_t とした時，LRR と BAF データをがん部の A-allele のコピー数を $n_{A,i}$，B-allale のコピー数を $n_{B,i}$ として

$$r_i = \gamma \log_2 \left\{ \frac{2(1-\rho) + \rho(n_{A,i} + n_{B,i})}{\Psi} \right\}, \quad (1)$$

$$b_i = \frac{1 - \rho + \rho n_{B,i}}{2 - 2\rho + \rho(n_{A,i} + n_{B,i})} \quad (2)$$

とモデル化する．ここで $\Psi = 2(1-\rho) + \rho\Psi_t$ であり，γ は測定機系によって決まるパラメーターである ($\gamma \sim 0.55$)．尚，式 (1)，(2) は正常細胞のみであればそれぞれ

$$\begin{aligned} r_i &= \gamma \log_2 \left(\frac{n_{A,i} + n_{B,i}}{2} \right), \\ b_i &= \frac{n_{B,i}}{n_{A,i} + n_{B,i}} \end{aligned}$$

と表される．ここで germline での CNV は無視した．実際には実データから得られる r_i 及び b_i，セグメント情報を使って

$$\begin{aligned} \hat{n}_{A,i} &= \frac{\rho - 1 + 2^{r_i/\gamma}(1-b_i)(2 - 2\rho + \rho\Psi_t)}{\rho}, \\ \hat{n}_{B,i} &= \frac{\rho - 1 + 2^{r_i/\gamma} b_i (2 - 2\rho + \rho\Psi_t)}{\rho} \end{aligned}$$

と予測する．

入力の A-allele 及び B-allele は Illumina が便宜的に決めた allele であり，各個人の haplotype はランダムに A-allele と B-allele の組み合わせになるが，ASCAT の出力ではセグメント毎に連続するコピー数として赤線と緑線で表示される (赤線 $\geq$ 緑線) 様である．またがん細胞のみのデータを入力することも可能で，その場合には ASCAT に登録されているチップ情報を指定して実行する．

2.1.3 Liu *et al.* (2010)[3] (MixHMM)

非がん部の細胞が混入した状態のがん部の Illumina の log R ratio (LRR) と B allele frequency (BAF) から，7 コピーまでの 20 state の hidden Markov model (HMM) を用いて Viterbi アルゴリズムで状態を決定

する方法論である．特に germline のデータは利用していない．

ある状態 i から他の状態に遷移する HMM の transition probability ρ_i は，π_i を状態 i の分布確率 (ここでは SNPs の割合) とすると，

$$\rho_i \equiv (1-\pi_i)(1-e^{-d/L_i})$$

と定義する．ここで d は 2 つの SNPs 間の距離 (塩基数) であり，$L_i = \lambda_i(1-\pi_i)$ (λ_i は状態 i が続く平均長) である．この ρ_i を用いて状態 i から状態 j への遷移行列 A は，

$$A_{ij} = \begin{cases} 1-\rho_i, & \text{if } i=j, \\ \frac{\rho_i \pi_j}{\sum_{l\neq i}\pi_l}, & \text{if } i \neq j \end{cases}$$

とする．一方 emission probability は，LRR と BAF で独立であると仮定し，状態 i から LRR の測定値 r が出力される確率は，

$$\Pr(r|i) \equiv \tau\varepsilon(r) + \frac{(1-\tau)}{\sigma_{i,R}}\phi\left(\frac{r-\mu_{i,R}}{\sigma_{i,R}}\right)$$

と定義する．ここで τ は全体的な fluctuation が起こる確率 (default 0.01) を表し，$\varepsilon(r)$ は r 値に依存する一定の確率密度関数，ϕ は正規分布であり，$\mu_{i,R}$ 及び $\sigma_{i,R}$ は状態 i の LRR の平均と分散である．同様に BAF の emission probability は，

$$\begin{aligned}\Pr(b|i) \equiv{}& \tau\varepsilon(b) + (1-\tau) \\ &\times \left[p_A^2 f(b;\mu_{i,o_A},\sigma^2_{i,o_A}) + p_B^2 f(b;\mu_{i,o_B},\sigma^2_{i,o_B})\right. \\ &\left. + p_A p_B\{f(b;\mu_{i,e_A},\sigma^2_{i,e_A}) + f(b;\mu_{i,e_B},\sigma^2_{i,e_B})\}\right]\end{aligned}$$

と定義する．ここで o_A，o_B は germline でそれぞれ AA，BB 由来，e_A，e_B は germline で AB 由来の遺伝子型クラスを表し，p_A 及び p_B は集団中の allele 頻度であり，μ 及び σ はそれぞれの状態の平均と分散である．確率密度 f は，

$$f(b;\mu,\sigma^2) = \begin{cases} \psi\left(\frac{-\mu}{\sigma}\right) & \text{if } b=0, \\ 1-\psi\left(\frac{1-\mu}{\sigma}\right) & \text{if } b=1, \\ \frac{1}{\sigma}\phi\left(\frac{b-\mu}{\sigma}\right) & \text{if otherwise} \end{cases}$$

としている．ここで ψ は正規分布の累積分布関数である．

正常細胞との混合については，正常が含まれる割合を p として LRR の値を

$$R_M = pR_N + (1-p)R_M$$

と表し，正常細胞からの寄与 R_N とがん細胞からの寄与 R_M がそれぞれ独立に log-normal distribution に従うとしてモデル化する．同様に遺伝子型 g の BAF を

$$b_{g,M} = w_N b_{g,N} + w_T b_{g,T}$$

とモデル化する．ここで $w_N = pn_N/(pn_N+(1-p)n_T)$，$w_T = (1-p)n_T/(pn_N+(1-p)n_T)$，$n_N$ 及び n_T はそれぞれ正常及びがん部でのコピー数である．更に $b_{g,N}$ 及び $b_{g,T}$ は正規分布に従うと仮定する．

実際の入力データは，BAF が 0.5 から対称になるように調整 [5] し，更に LRR は GC 含有量による補正 [1] を行い，更に目視によって 1 copy もしくは 1 copy 領域を指定した上でソフトウエアを実行する．

2.1.4 Li *et al.* (2011)[2] (GPHMM)

Global parameter hidden Markov model (GPHMM) はがん細胞の aneuploidy，正常細胞の混合，GC 含有量による bias を考慮したコピー数推定ソフトウエアである．入力データはがん部と非がん部の Illumina の log R ratio (LRR) と B allele frequency (BAF) である．

HMM の state は 5 コピーまでに対応した 12 状態であり，signal fluctuation を指定する特別な状態"0" (がん細胞の両 loss と同様の状態だが陽動項として区別している) を加えている．GPHMM では，正常細胞の割合 w_s，LRR baseline shift o，GC 含有量の重み h，LRR 及び BAF の標準偏差 σ_l，σ_b の 5 つの global parameter を用いてモデル化している．HMM の状態が c の時の i 番目の SNP の LRR が l_i となる確率密度を

$$f(l_i|w_s,o,h,\sigma_l,c) = \frac{1}{\sigma_l}\phi\left(\frac{l_i - \{2\log_{10}\frac{y_c}{2} + o + hg_i\}}{\sigma_l}\right)$$

とする．ここで $\phi(x)$ は正規分布，g_i はその SNP を検出するプローブの GC 含有量である．y_c は平均コピー数で $y_c = w_s n_s + (1-w_s)n_c$，$n_s$ 及び n_c はそれぞれ正常及びがん細胞でのコピー数である．同様に BAF b_i は，

$$f(b_i|w_s,\sigma_b,c) = \sum_{k=1}^{g_c}\frac{p_0(k)}{\sigma_b}\phi\left(\frac{b_i - \frac{w_s n_s u_{sk} + (1-w_s)n_c u_{ck}}{y_c}}{\sigma_b}\right)$$

とモデル化する．ここで g_c は状態 c における遺伝子型の種類の数であり，$p_0(k)$ は遺伝子型 k を観測する事前確率，u_{sk} 及び u_{ck} はぞれぞれ正常細胞及び純粋ながん細胞における BAF の理論値である．5 つの global parameter を EM algorithm でデータから決定後に HMM によってコピー数の推定を行う．

ソフトウエアはMatlab/C上で動作し，各SNPsのGC含有量やBAFはPennCNV[7]のパッケージに含まれている情報を用いる．

2.2 Array comparative genomic hybridization データの補正

2.2.1 Diskin *et al.* (2008)[1)]

Illumina及びAffymetrixのどちらのSNP genotyping arrayのデータについてもゲノム上でのSNP位置に依存したハイブリダイゼーション測定値の変動がみられており，筆者らはこれを「genomic wave」と呼んでいる．本論文ではgenomic waveに起因するゲノム上の局所位置でのデータ変動の確認，ゲノム上の局所的特徴の相関解析を行い，補正方法について報告を行っている．

Illumina (HumanHap1M array, HumanHap550 array) 及びAffymetrix (Mapping 500K array, genome-wide 6.0 array) を用い，何れのarrayにもlog R ratio (LRR) にgenomic waveが発生することが確認された．更にゲノム上の局所的特徴として，GC含有量 (%)，segmental duplication，遺伝子含量，エクソン含量，単純反復，遺伝子保存領域との相関解析を行っている．解析の結果$1Mb$窓でのGC含有量 (%) とLRRとの相関係数が高いことが見いだされた．

またIllumina HumanHap550 arrayを用いて測定されたLRR値についてGC含有量 (%) がポジティブに働くか，ネガティブに働くかについても確認を行っている．測定の際の総DNA量が推薦値である750ngより少ない場合にはポジティブに働き，750ngより多い場合にはネガティブに働くという結果を得ている．

Genomic waveの影響を考慮してLRR値を補正するためM個の全マーカーから互いに少なくとも$1Mb$以上離れたm個のマーカーを選択して解析に用いている．実際の補正ではまずm個のマーカーのLRR値$L_j(j=1,\ldots,m)$，マーカーの周辺の窓1 Mbでの GC含有量 (%) により

$$L_j = \alpha + \beta \times G_j + \epsilon_j$$

と回帰モデルを作成する．ここでαとβはモデルパラメーターであり，最小二乗法により推定する．ただし実際に生じているCNVの影響を避けるためLRR値は$[-2,1]$の範囲のものを使用している．

モデルパラメーターを求めた後，全マーカーに対して周辺の窓1 MbのGC含有量に基づいて期待されるLRR値を算出する．観測LRR値から期待されるLRR値を引いた値 (残差) が補正されたLRR値となる．

補正の手続きはPennCNVパッケージ[1]に含まれている．スタンドアロンとして実行できるプログラム及びPennCNVの内部に組み込まれたプログラムとして使用可能である．

2.2.2 Staaf *et al.* (2008)[5)](tQN)

Illumina genotyping arrayの測定において蛍光のdyeの違いに起因すると考えられるアリル強度分布の違いに着目して考案したアリル強度の補正方法 (tQN) について報告している．

Illuminaでは任意の遺伝子座の対立遺伝子をA alleleおよびB alleleと呼んでおり，規格化された蛍光強度をそれぞれX，Yとしている．これらはラベリングに使用する蛍光ダイ (Cy5およびCy3) の違いに対応したものである．規格化された蛍光強度値X，Yは，BeadStudio(現GenomeStudio) により実際に測定される蛍光強度から比較可能なアリル強度の指標となるような変換が施される．しかしながら全ゲノム上のSNPsに対するアリル強度の分布には違いが残っており，この違いをquantile normalizationによって補正する．ただし分布の特性上Xが小さいものについて補正が大きすぎる場合があり，それを防ぐため補正後の値に閾値を設定している．元の値X，Yに対して補正値が1.5倍を超える場合には，$1.5*X$，$1.5*Y$としている．文献[5)]で示されていた具体例を図1に転載した．

各SNPsのLRRやBAFを算出するには補正されたクラスター中心が必要であるため，HapMapサンプル (Illumina 300k version 1: n=111, version 2: n=120, 370k: n=123, 550k: n=120) を用いてreference data setを作成している．

tQNはRとPerlで記述されており，tQN Project web pageからダウンロード可能である[2]．プログラムの他に，Illumina Human660W-Quad, Human1M-omnia, HumanOmni2.5-Quad, HumanOmniExpressの参照データ (cluster file) がダウンロード可能である．

2.3 CNA推定ソフトウエアのまとめ

調査を行ったCNA推定ソフトウエアの概要を表2にまとめた．

[1] http://www.openbioinformatics.org/penncnv
[2] http://baseplugins.thep.lu.se/wiki/se.lu.onk.IlluminaSNPNormalization

表 2 調査対象 CNA 推定ソフトウエアのまとめ. LRR:log R ratio，BAF:B allele frequency，PFB:population frequency of B allele，HMM:hidden Markov model.

	GenoCNA	ASCAT	MixHMM	GPHMM
前処理	Wave 補正	Wave 補正	Wave 補正 (+tQN)	無し
入力データ	LRR，BAF，PFB	LRR，BAF	LRR，BAF，PFB，1 コピー及び LOH 領域	LRR，BAF，PFB
アルゴリズム	HMM	モデル fitting	HMM	HMM
対象範囲	常染色体，X 染色体	常染色体，X 染色体	常染色体	常染色体
非がん部の利用	遺伝子型情報の抽出	ヘテロの抽出	利用無し	利用無し
出力情報	両アリル数，事後確率	混入率，ploidy，両アリル数，推定精度	両アリル数	混入率，両アリル数，事後確率

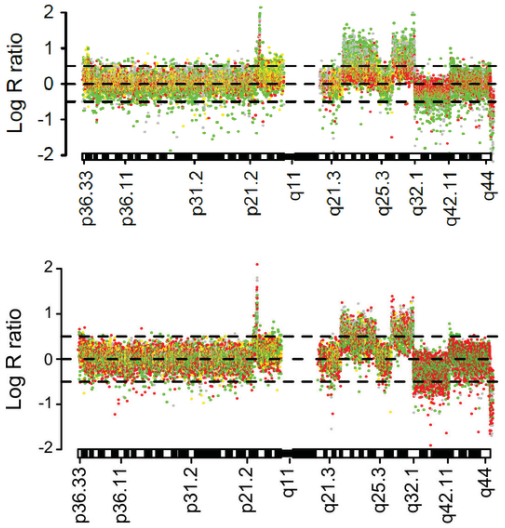

図 1 コピー数推定における tQN の効果 (文献 [5] の Figure 1 (d) 及び Figure 5 (c))。Urothelial tumor サンプルの 1 番染色体の Log R ratio を 遺伝型で色分けしたもの (AA:緑，AB:黄，BB:赤). 上段は Illumina の BeadStudio によるもの，下段は tQN 実施後のもの. BeadStudio では，AA(緑) と BB(赤) がオーバーラップしておらず非対称であるが，tQN 適用後は改善されている.

3 Omni1 データによるソフトウエアの評価

国立がん研究センターより提供された Omni1 の固形がんの実データを用いて，GenoCNA[6]，ASCAT[4]，MixHMM[3]，GPHMM[2] の評価を行った.

3.0.1 CNA の推定手順

下記に示す手順でそれぞれのソフトウエアを実行した.

3.0.2 GenoCNA[6]

1. がん部，非がん部の LRR に対して genomic wave[1] 補正を実行.
2. 非がん部のデータ (LRR，BAF) と，B アリル集団頻度 (PFB) を用い，R 上にて genoCNV を実行して，非がん部の各 SNP の遺伝子型 (B アリルコピー数) を推定.
3. がん部のデータ (LRR，BAF)，B アリル集団頻度 (PFB)，及び推定した遺伝子型 (B アリルコピー数) を用い，R 上にて genoCNA を実行.
4. genoCNA の実行結果 (HMM の状態によるセグメント) に，コピー数非変化領域 (genoCNA の segment 形式の出力では CNA 領域のみ出力するため)，及びセントロメアの情報を付加.

3.0.3 ASCAT[4]

1. がん部，非がん部の LRR に対して genomic wave[1] 補正を実行.
2. がん部，非がん部別にそれぞれ LRR，BAF 毎にまとめたファイル (ASCAT の入力形式) を作成し，R 上で ASCAT を実行.
3. R 上の ASCAT の実行結果からセグメント毎の出力をテキストファイルに出力.
4. 出力されたセグメントデータにセントロメアの情報を付加.

3.0.4 MixHMM[3)]

1. がん部の LRR に対して genomic wave[1)] 補正を実行 (tQN は実行せず).
2. がん部のデータ (LRR，BAF) について，国立がん研究センターが目視によって 1 コピー，2 コピー領域を特定.
3. 上記に基づき，がん部のデータ (LRR，BAF) について，1 コピー領域 (がん部 A，非がん部 AB 混合) での BAF 値の median，及び 2 コピー領域での LRR 値の median を算出.
4. がん部のデータ (LRR，BAF)，B アリル集団頻度 (PFB)，及び上記 LRR 値，BAF 値を用い MixHMM を実行.
5. MixHMM の実行結果 (HMM の状態によるセグメント) にセントロメアの情報を付加.

3.0.5 GPHMM[2)]

1. がん部のデータ (LRR，BAF) と，PFB(B アリル集団頻度)，各 SNP 位置周辺の ±500 kb の GC 含有量を用い，Windows (64bit) 上で GPHMM を実行 (Genomic Wave 補正は GPHMM により行われる).
2. GPHMM の実行結果 (HMM の状態によるセグメント) にセントロメアの情報を付加.

3.1 CNA の推定結果の比較

サンプル毎の推定結果を図 2，3 に示した．ASCAT では 5 サンプルで推定が失敗して 95 サンプル，MixHMM では 3 サンプルで 1 copy 領域が目視で特定出来なったために 97 サンプルの結果となっている．また図 4 に ASCAT 及び GPHMM で推定されたがん細胞の割合と領域長で平均したコピー数 (ploidy) を示した.

GenoCN は 2 copy を基準に推定している様で，ploidy の増加への対応が不十分である様に見える．ASCAT は細かい CNA が多数出力され，また積極的に ploidy の増加を採用している様である．MixHMM は 1 copy 若しくは 2 copy の領域を人が決めて入力する必要があり，この入力の精度により大きく結果が変わるが，ホモとなる SNPs が続くと LOH としてしまう傾向が強い様に思われる．GPHMM も同様に LOH と予測する領域が多く，また loss 領域において 1 copy と 2 copy の LOH を細かくモザイク状に予測することがある様である．これらは ASCAT と同様に，非がん部でヘテロとなった SNPs のみを MixHMM や GPHMM の入力にすると改善するのかもしれないと考え，GPHMM の入力を非がん部でヘテロになった SNPs に限定してみることにした．BK0001D と BK0002D に対して推定を行い，他の結果と比較した図を図 5 に示した．BK0001D で見られた 2 copy での LOH のモザイクが消えており，BK0002D でも概ねモザイクが消えている．一方で BK0002D の 3p においては loss ではなく copy neutral LOH と予測しており，また 5q では 4 ∼ 2 copy でモザイク状に推定が揺いでしまった．GPHMM においてヘテロのみを入力にすると改善される領域も多いが，逆に推定が不安定になってしまう領域もある様である.

図 2，3 からは，これらのソフトウエアによる予測では多数の細かい CNA 領域が出力されることがあることが分かる．これはソフトウエア開発に使用したデータの SNPs 数が Omni-1 に比べて少ないことや，日本人であるために allele 頻度が異ること等が原因として考えられるのかもしれない．そこで主観的ではあるが多数の細かい CNA は偽陽性の方が多いと考えて，sensitivity は犠牲になるが 100Kbp 未満の領域をコピー数の状態が近い近接領域にマージした ASCAT と GPHMM の結果を図 6 に示した．細かな点が消えていることが分かる．参考に図 7 にサンプル毎に各予測ソフトウエアで出力された常染色体のセグメント数を示した (GPHMM のみ X 染色体をサポートしていない)．上述の様に MixHMM ではモザイク状に LOH と予測することがあるため，セグメント数の突出が散見される．セグメント数はソフトウエアによっては大きく異り，GenoCNA や GPHMM でも極端に多いサンプルが存在する一方で，ASCAT による予測は総じてやや多い様に見受けられる．明確な根拠は無いが 100Kbp 未満の領域を近接領域にマージしてしまうと，ASCAT と GPHMM のセグメント数は相当数抑えられている (表 3 参照).

上記の予測結果それぞれに対して，loss 及び LOH のヒストグラムを図 8 に，gain のヒストグラムを図 9 に示した．ヒストグラムにすると形状はどれも似ているようである．ASCAT と GPHMM の形状が特に似ているが，ASCAT の方がやや gain と予測するサンプルが多い様である．また smoothing を施すと揺らぎが減少するためヒストグラムが明瞭になり，セントロメア付近のピークも一部小さくなっている.

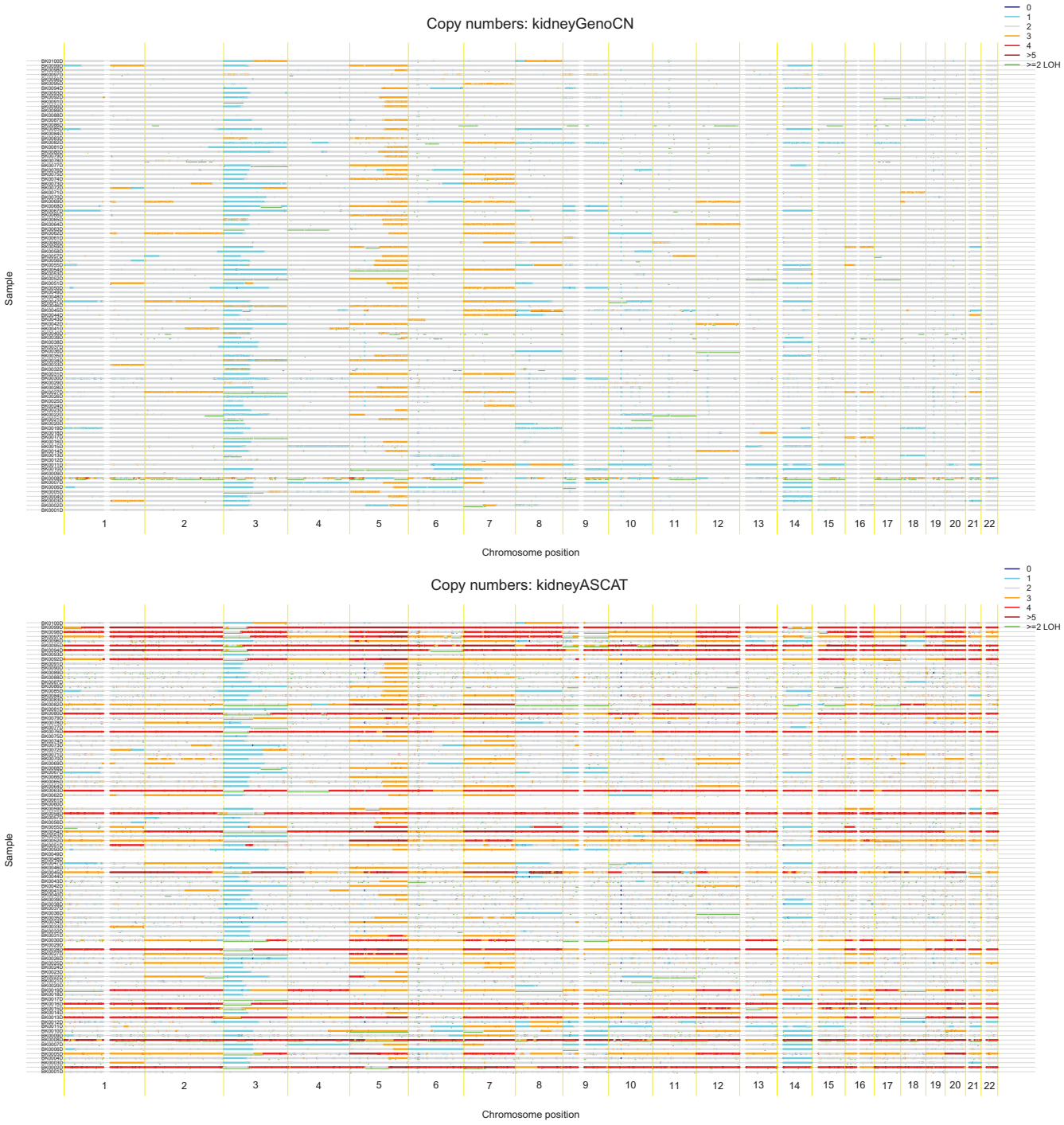

図 2　Wave 補正後の固形がんの Omni1 のデータを genoCN(上段)，ASCAT(下段) で推定した結果.

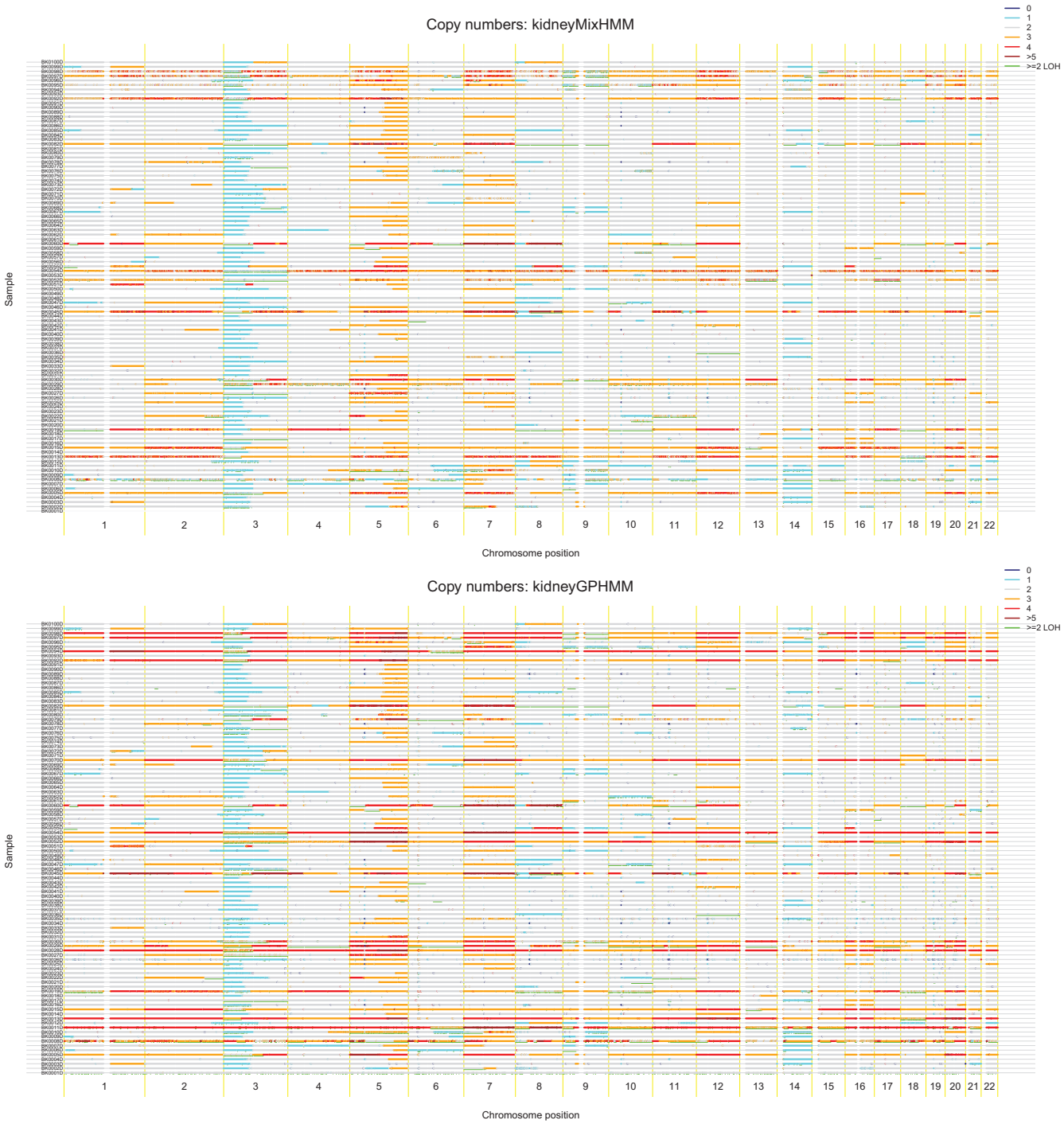

図 3 Wave 補正後の固形がんの Omni1 のデータを MixHMM(上段) と，wave 補正無しのデータから実行した GPHMM(下段) で推定した結果.

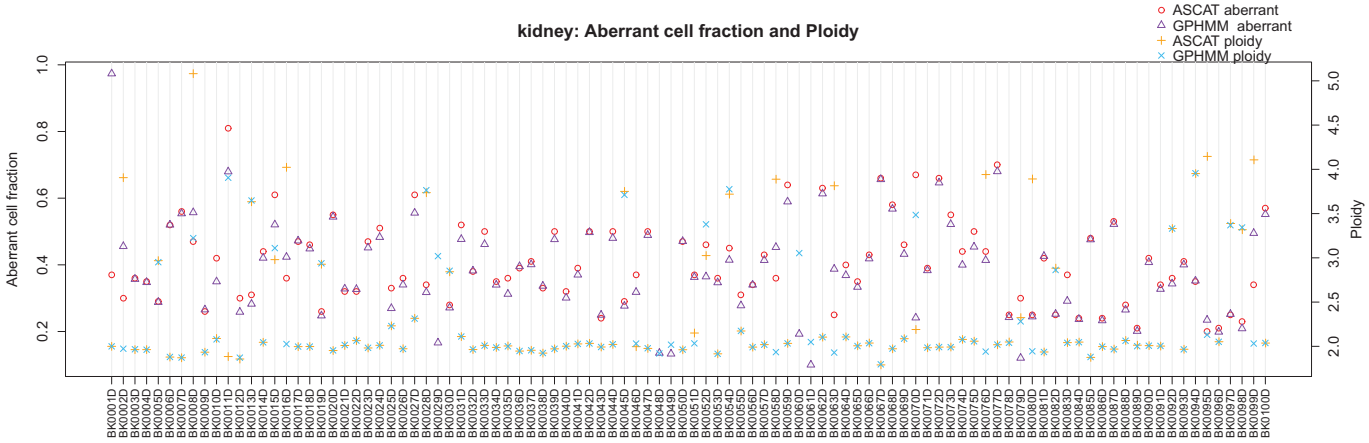

図 4　ASCAT 及び GPHMM で予測されたがん細胞の割合 (左縦軸) と平均コピー数 (右縦軸:ploidy).

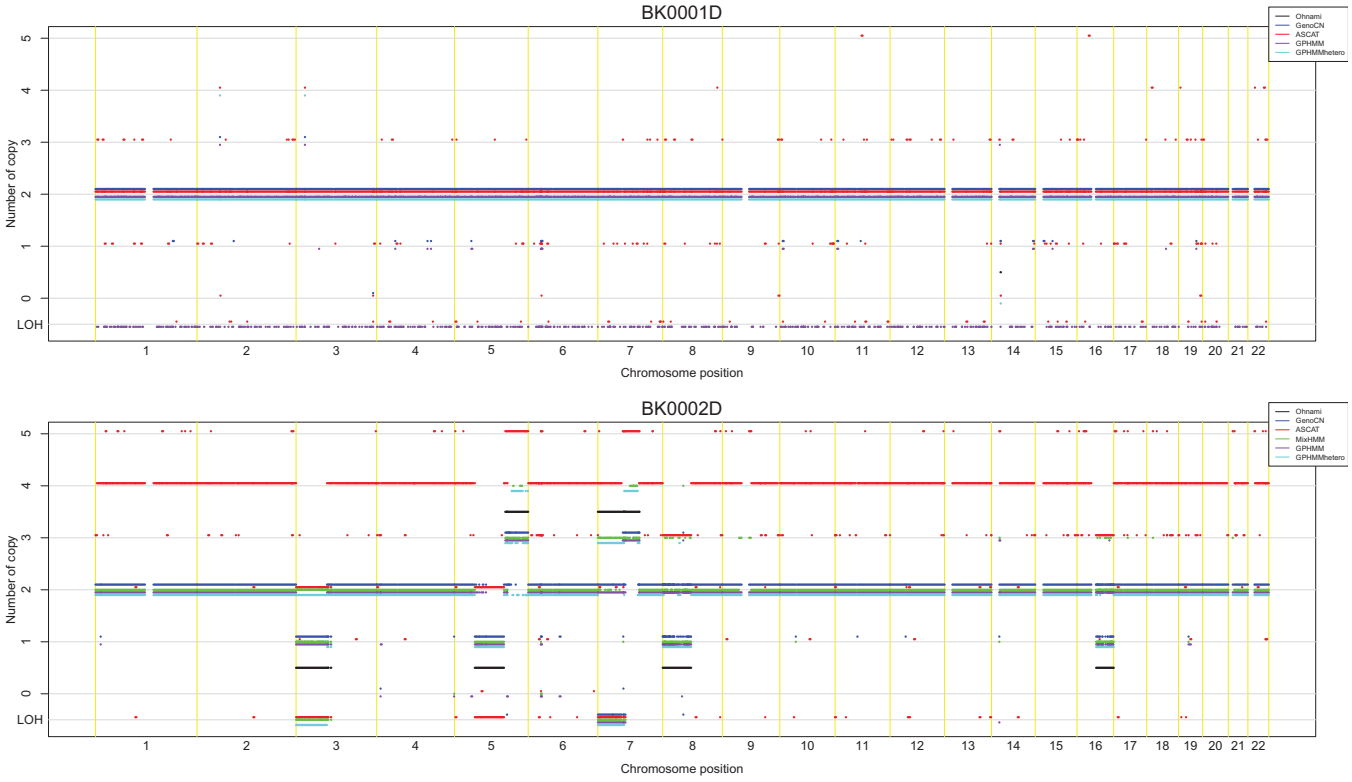

図 5　GPHMM の入力に全ての SNPs を使った場合と非がん部でヘテロになった SNPs のみを使った場合の比較.

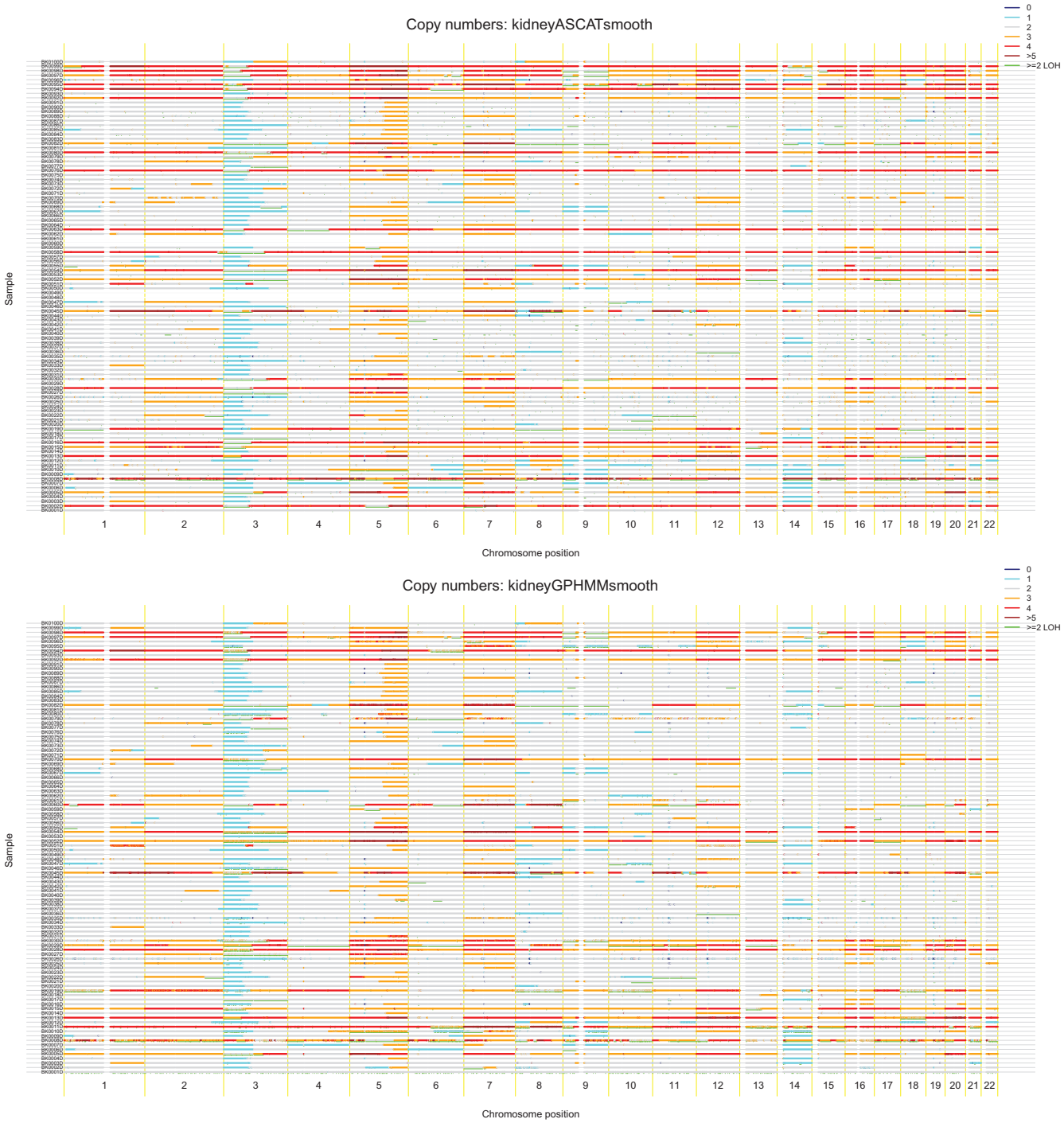

図 6　上段:ASCAT の推定結果から 100Kbp 未満の領域を統合した結果，下段:GPHMM の推定結果から 100Kbp 未満の領域を統合した結果.

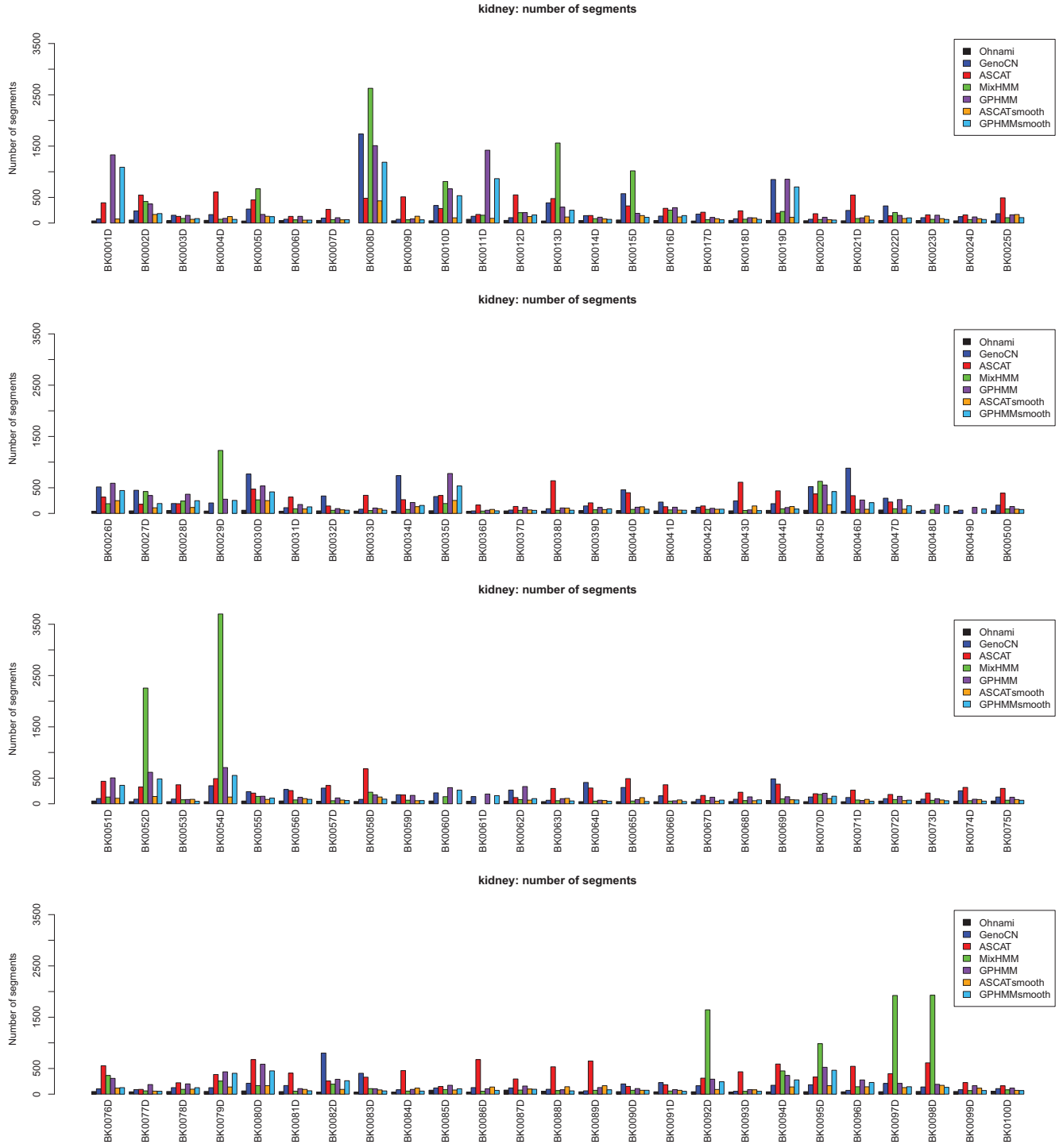

図 7 Ohnami，GenoCN，ASCAT，MixHMM，GPHMM，ASCAT smooth，GPHMM smooth によるサンプル毎の常染色体のセグメント数.

表 3 予測結果の総セグメント数の一覧. 染色体の短腕，長腕毎にカウント.

Method	Ohnami	GenoCN	ASCAT	MixHMM	GPHMM	ASCAT smooth	GPHMM smooth
#Sample	99	100	95	97	100	95	100
Number	5,077	28,926	32,182	30,829	25,439	10,940	17,937

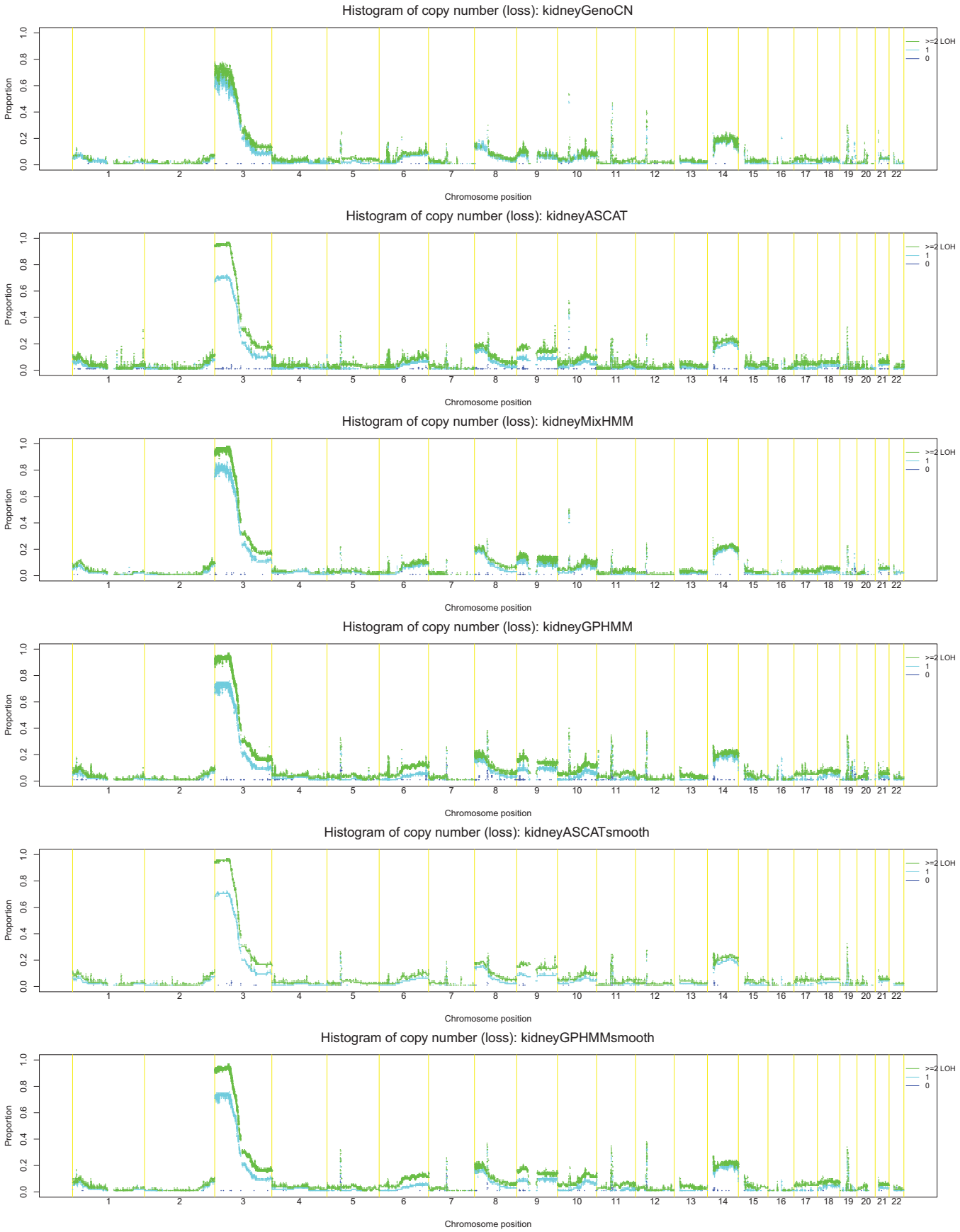

図 8　上段から GenoCN，ASCAT，MixHMM，GPHMM，ASCAT smooth，GPHMM smooth による loss 及び LOH の予測結果のヒストグラム.

図 9　上段から GenoCN，ASCAT，MixHMM，GPHMM，ASCAT smooth，GPHMM smooth による gain の予測結果のヒストグラム.

3.2 CNA の推定のまとめ

固形がんの Omni1 データを用いて，GenoCNA[6]，ASCAT[4]，MixHMM[3]，GPHMM[2] の評価を行った．これらの結果をまとめると，

1. GenoCNA では ploidy の増幅に対応出来ない．
2. ASCAT は予測に失敗するサンプルが無視出来ない数存在する．
3. ASCAT は細かな CNA を積極的に予測し，また ploidy の増幅も他よりは積極的に採用する．
4. MixHMM は目視によって 2 copy 及び 1 copy(若しくは LOH) 領域の同定が必要であり，この結果に推定が大きく依存する．
5. MixHMM は細かなモザイク状に LOH を予測し，非常に多くのセグメントに分割してしまうことがある．
6. GPHMM は MixHMM と同様にモザイク状に LOH を予測することがある (MixHMM よりは少ない様である．GPHMM は MixHMM の後継ソフトウエアであり，改善された箇所かもしれない)．
7. ソフトウエアによる予測は，小さなセグメントを多数出力することがある．

と言えるだろう．そこで CNA 予測方法として，

1. ASCAT と GPHMM で予測．
2. 必要に応じて細かな CNA を近接領域に繋げてしまう．
3. 上記 2 つの予測結果を提示し，予測がずれている箇所等は別途実験的な検証を必要とする候補領域とする．

を提案する．ASCAT と GPHMM の予測性能をこれ以上比較するには，実験的に確定したデータが必要であり，これらが無い現状では 2 つのソフトウエアでの予測結果を提示することで対応したい．

4 Illumina SNPs チップデータからの CNA 予測パイプライン

3 節の結果から，Illumina SNPs チップデータからの CNA 予測パイプラインを

1. ASCAT による予測
   (a) がん部，非がん部の LRR に対して genomic wave[1] 補正を実行
   (b) 上記データに対して ASCAT を実行
   (c) ASCAT の予測結果の内，小さい領域を周辺領域にマージ
2. GPHMM による予測
   (a) がん部のデータ (genomic wave 補正無し) と PFB(B アリル集団頻度)，各 SNP 位置周辺の ±500 kb の GC 含有量データを用いて GPHMM を実行
   (b) GPHMM の予測結果の内，小さい領域を周辺領域にマージ
3. 上記 2 つの予測結果を染色体上に表示し，比較

とした．本パイプラインによる ASCAT 及び GPHMM の予測結果をそれぞれクラスタリングした結果を図 10 及び図 11 に示した (loss と gain のヒストグラムは図 8，9 の再掲)．

謝辞：本稿は，独立行政法人 医薬基盤研究所の保健医療分野における基礎研究推進事業にて実施された「多層的オミックス解析による創薬標的の網羅的探索を目指した研究」の「多層的疾患オミックス解析における，ゲノム情報に基づく創薬標的の網羅的探索を目指した研究」において実施された調査の一部である．弊社技報への転載をご許可頂いた共著者各位に感謝申し上げる．

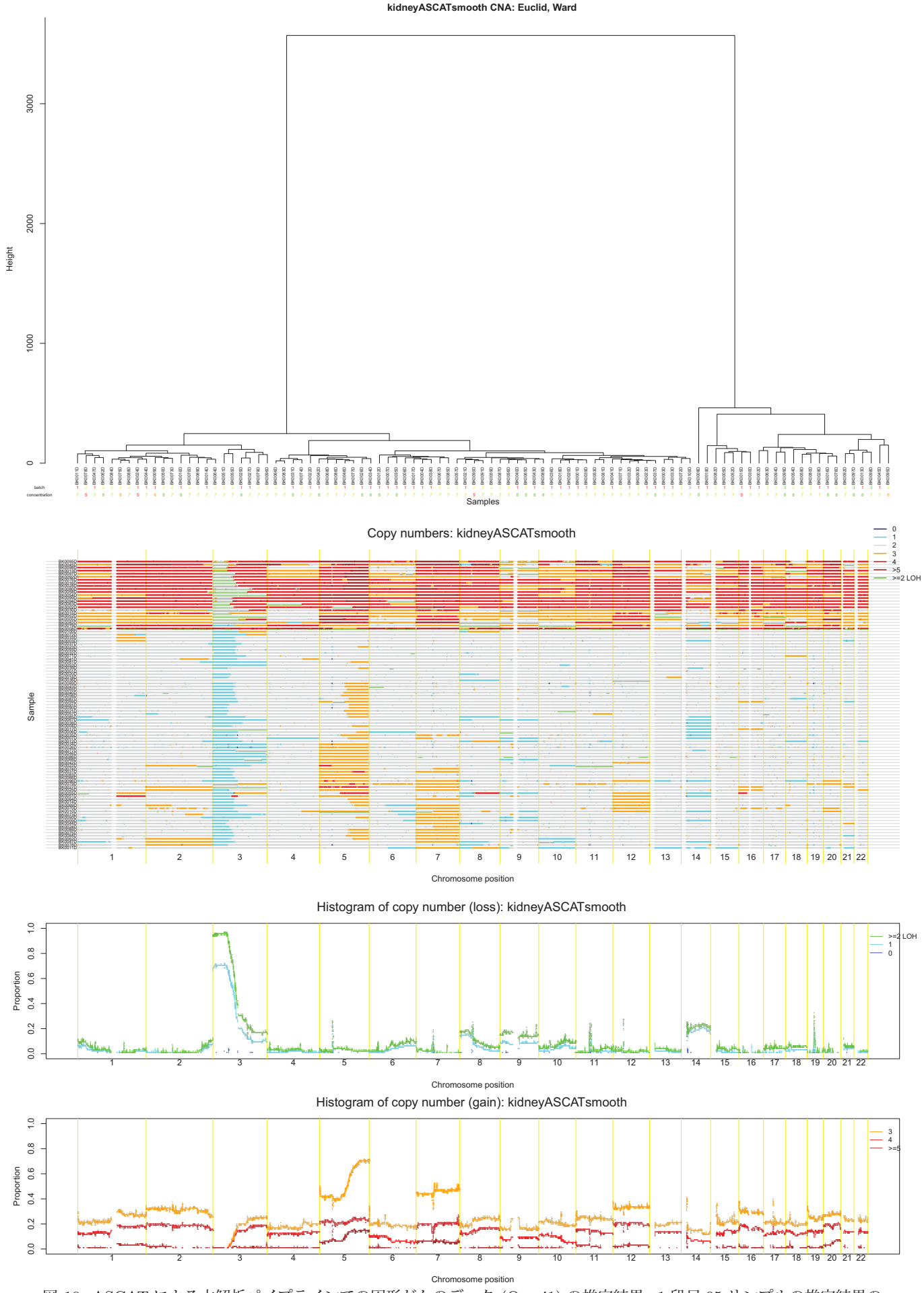

図 10　ASCAT による本解析パイプラインでの固形がんのデータ (Omni1) の推定結果．1 段目:95 サンプルの推定結果のクラスタリング (Euclid 距離，Ward 法)，2 段目:各サンプルのコピー数をクラスタリング順に表示，3 段目:loss のヒストグラム，4 段目:gain のヒストグラム．

図 11　GPHMM による本解析パイプラインでの固形がんのデータ (Omni1) の推定結果．1 段目:100 サンプルの推定結果のクラスタリング (Euclid 距離，Ward 法)，2 段目:各サンプルのコピー数をクラスタリング順に表示，3 段目:loss のヒストグラム，4 段目:gain のヒストグラム．

引 用 文 献

1) Sharon J. Diskin, Mingyao Li, Cuiping Hou, Shuzhang Yang, Joseph Glessner, Hakon Hakonarson, Maja Bucan, John M. Maris, and Kai Wang. Adjustment of genomic waves in signal intensities from whole-genome snp genotyping platforms. *Nucleic Acids Research*, Vol. 36, No. 19, p. e126, 2008.

2) Ao Li, Zongzhi Liu, Kimberly Lezon-Geyda, Sudipa Sarkar, Donald Lannin, Vincent Schulz, Ian Krop, Eric Winer, Lyndsay Harris, and David Tuck. Gphmm: an integrated hidden markov model for identification of copy number alteration and loss of heterozygosity in complex tumor samples using whole genome snp arrays. *Nucleic Acids Research*, Vol. 39, No. 12, pp. 4928–4941, 2011.

3) Zongzhi Liu, Ao Li, Vincent Schulz, Min Chen, and David Tuck. Mixhmm: Inferring copy number variation and allelic imbalance using snp arrays and tumor samples mixed with stromal cells. *PLoS ONE*, Vol. 5, No. 6, p. e10909, 06 2010.

4) Van Loo P, Nordgard SH, Lingjaerde OC, Russnes HG, Rye IH, Sun W, Weigman VJ, Marynen P, Zetterberg A, Naume B, Perou CM, Borresen-Dale AL, and Kristensen VN. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci USA*, Vol. 107, No. 39, pp. 16910–16915, September 2010.

5) Johan Staaf, Johan Vallon-Christersson, David Lindgren, Gunnar Juliusson, Richard Rosenquist, Mattias Hoglund, Ake Borg, and Markus Ringner. Normalization of illumina infinium whole-genome snp data improves copy number estimates and allelic intensity ratios. *BMC Bioinformatics*, Vol. 9, No. 1, p. 409, 2008.

6) Wei Sun, Fred A. Wright, Zhengzheng Tang, Silje H. Nordgard, Peter Van Loo, Tianwei Yu, Vessela N. Kristensen, and Charles M. Perou. Integrated study of copy number states and genotype calls using high-density snp arrays. *Nucleic Acids Research*, Vol. 37, No. 16, pp. 5365–5377, 2009.

7) Kai Wang, Mingyao Li, Dexter Hadley, Rui Liu, Joseph Glessner, Struan F.A. Grant, Hakon Hakonarson, and Maja Bucan. Penncnv: An integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data. *Genome Res.*, Vol. 17, No. 11, pp. 1665–1674, 2007.