全エクソンシーケンシングデータからの生殖細胞系ゲノムコピー数予測ソフトウエアの調査

佐藤智之[i] 知久季倫[i] 坂本裕美[ii] 吉田輝彦[iii] 岩崎基[iv] 津金昌一郎[v]

A research of germline DNA copy number variation prediction sotfware from whole exome sequencing data

Toshiyuki Sato Suenori Chiku Hiromi Sakamoto
Teruhiko Yoshida Motoki Iwasaki Shoichiro Tsugane

Whole exome sequencing (WES) データから germline の copy number variataion (CNV) を予測するツールの文献を調査し，幾つかのソフトウエアについて whole genome sequencing 及び SNPs チップからの CNV 予測結果と比較した．この結果感度及び陽性的中率はかなり低く，WES データから CNV を予測することの困難さが浮彫りになった．

(**キーワード**): DNA, CNV, NGS, WES

1 はじめに

ヒトゲノム上には copy number variataion (CNV) と呼ばれる比較的大きな (1 kbp 以上) 挿入，欠損が広く存在することが知られている [10]．CNV は様々な疾患との関係も研究されている生殖細胞系の多型 (copy number polymorphism; CNP) の一つであり，ゲノム上における網羅的な検出は genome wide association study (GWAS) の精度を高めることが期待される．そこで本調査では，next generation sequencing (NGS) の一つである whole exome sequencing (WES) データ (タンパク質をコードしている全ゲノム領域) からの CNV 検出ツールの調査を行い，この性能を評価した．尚，本調査ではがんにおける somatic なコピー数の変化である copy number aberrations/alternations (CNA) を対象としたツールは調査対象外としたが，ExomeCNV[32] は他で比較に用いられているため調査対象に含めた．

2 文献調査

2.1 Love *et al.* (2011)[25] (exomeCopy)

Consensus coding sequence (CCDS) について，x bp の CCDS を $\max(1, \lfloor x/100 \rfloor)$ 個の重ならない window に分けて (例えば 402 bp の CCDS であれば 4 個の window) read count データを作成して CNV を予測する (exon 単位ではない)．CNV call はコピー数の状態を $S_i = \{0, 1, 2, 3, 4\}$ (男性の X 染色体については $S_i = \{0, 1, 2\}$) とした HMM を採用し，各状態からの read count の出力確率は，平均 μ を background read detph，window 幅，GC 含有量から線形回帰で求めた値，dispersion パラメーター ϕ を $\hat{\phi} = \max\{(s^2 - \bar{o})/\bar{o}^2, \varepsilon\}$ とした負の 2 項分布であるとし，遷移確率は window 間の距離の関数とした．ここで $\bar{o}$ はサンプルにおける window 毎の read count の平均値，s^2 は read count の線形回帰から分散であり，ε はモデルにおける ϕ の下限値 (計算機における 0 で無い最小の正の数) である．尚，ϕ を background SD 及び background variance から線形回帰で求めた値に変更した予測法を exomeCopyVar と呼んでいる．HMM の状態の予測は Viterbi アルゴリズムを採用している．

実データへの適用として，男性 248 人の X-linked Intellectual Disabilities (XLName) プロジェクトのデータを exompeCopy で解析している．XLName ではカスタム Agilent SureSelect (target 長 3.8 Mb) を用いて 76 bp の single read データを sequencing し，RazerS[36] でゲノムにマッピングしている．このデータから 11,581 CNV を予測し，この内 5 window 以上で予測された CNV は 640 箇所であった．10 kb 未満の CNV は DGV に登録がある物も多いが，100kb 以上になると少なくなる．またバックグラウンドとして使うデータについてはある程度頑健であり，NimbleGen array を使った

[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 遺伝医学研究分野 ユニット長
[iii]国立がん研究センター 研究所 遺伝医学研究分野 分野長
[iv]国立がん研究センター がん予防・検診研究センター 疫学研究部 部長
[v]国立がん研究センター がん予防・検診研究センター センター長

データ (Danish exome data) でもバックグラウンドを使わないよりは良い成績であった.

Sensitivity を評価するために Danish exome data の 1 番染色体の read 数を調整することによりシミュレーションデータを作成し，aCGH 用に開発された DNAcopy[34] 及び BioHMM[24] と比較した．この結果 exomeCopy が常に同等以上の sensitivity を示し，更にシミュレーションデータに導入した CNV 以外は 0.4% しか call しなかったと報告している．また exomeCopy はバックグラウンドデータと比較しているため CNV の頻度が高いと検出力が落ちるが，頻度 25%程度の CNV でも 50%程度は検出出来たとしている．Read depth による予測能については，1000 人ゲノム [9] の PUR 集団のデータを用いてシミュレーションデータを作成し，window 毎の read count が 50 以上あれば 78%程度の CNV を検出することが出来た.

2.2 Sathirapongsasuti *et al.* (2011)[32] (ExomeCNV)

Exon 単位の read count (本文献で depth-of-coverage と記載しているが，本文中に read count を depth-of-coverage と言うと記載されている) と B-allele frequencies (non-reference call frequencies) を用いて，control サンプルに対する case サンプルの CNV と loss of heterozygosity (LOH) を別々に予測する．CNA 予測ツールであるが，後の文献では CNV 予測の比較で用いられているため紹介しておく.

入力は総 read 数で規格化後のがん部 (case) と非がん部 (control) の WES データとし，ある exon の read count をそれぞれ X，Y とする．これがそれぞれ独立に Poisson 分布に従うと仮定すれば，十分な read 数であれば平均と分散がそれぞれ λ_X 及び λ_Y の正規分布 $N(\lambda_X, \lambda_X)$，$N(\lambda_Y, \lambda_Y)$ で近似することが出来る．ここで $R = X/Y$，$\lambda_X = \rho\lambda_Y = \rho\lambda$ (deletion:$\rho = 0.5$, duplication:$\rho = 1.5$) とすると，R は平均や分散が定義出来ない Hinkley 分布に従ってしまうので，Greary-Hinkley 変換

$$t(\rho) = \frac{\lambda_Y R - \lambda_X}{\sqrt{\lambda_Y R^2 + \lambda_X}} = \frac{(R-\rho)\sqrt{\lambda}}{\sqrt{R^2+\rho}} \quad (1)$$

を行い，正規分布に従う $t(\rho)$ で CNV を検出する．$t(\rho)$ に対してカットオフ値 $r(\rho)$ を与えることにより，receiver operating characteristic (ROC) 曲線を描くことが出来る．非がん部細胞が混入している場合，がん細胞含有割合 c は LOH 解析から求めることが出来，$\rho' = c+\rho(1-c)$ と置き換えることによって同様に sensitivity と specificity を制御することが出来る．CNV 領域のセグメント化には，$\log R$ を用いて DNAcopy[34] の circular binary segmentation[27] (CBS) アルゴリズムを用いる． また LOH の推定では control サンプルで多型がある箇所 i における非リファレンス allele を B-allele と定義し，B_i を B-allele のカウント数，N_i を i における全 read カウントとすると，BAF は $\mathrm{BAF}_i = B_i/N_i$ と定義される．各セグメントにおいて case と control の BAF の分散を F 検定で評価することにより，LOH を推定する.

評価に使った実データは，Agilent SureSelect Human All Exon G3362 でキャプチャーした 76 bp の single-end で sequencing した皮膚 (非がん部) の 4 サンプルと melanoma 1 サンプル (depth 37.5x)，76 bp の paired-end の皮膚の 1 サンプルであり，Novoalign で hg18 にマッピングし，Picard で PCR duplicates を除いた．同じ皮膚サンプルを別々の 2 lane で流し ExomeCNV で評価したところ，異る call は得られず，specificity もほぼ理論値通りの結果が得られた．また男性を control として女性の性染色体を解析すると，X 染色体は duplicated と call され，Y 染色体は deleted と call された．続いて melanoma サンプルと matched pair となる皮膚サンプル (depth 42.8x) に対して ExomeCNV を実行し，これを Omni-1 のデータを用いて genoCNA[33] で CNA 推定を行った結果と比較した．また WES データであるが ERDS[40] でも解析を行い[1]，Omni-1 の結果を gold standard とした時の推定精度を合わせて表 1 に示した.

2.3 Plagnol *et al.* (2012)[29] (ExomeDepth)

WES データにおける exon 毎の read 数の over-dispersion を beta-binomial モデルで fit させ，コントロールセットと比較して HMM で CNV call を行う．コントロールセットに無い稀な CNV の検出に適しており，common CNV の検出には不向きである.

ある exon i におけるテストサンプルの read 数を X_i，統合した reference control の read 数を Y_i とした時に，X_i を

$$\pi_i \sim \mathrm{Beta}(\mu_i, \frac{\phi}{1-\phi}), \quad (2)$$

$$X_i \sim \mathrm{Bin}(p = \pi_i, n = X_i + Y_i) \quad (3)$$

$$(4)$$

[1] ERDS の web page には「WES データには適していない」とのコメントが強調されている.

表 1 Omni-1 のデータから genoCNA[33] で CNA 推定した結果を gold standard としたときの ExomeCNV と ERDS[40] の推定精度

	ExomeCNV			ERDS	
	Deletion	Duplication	LOH	Deletion	Duplication
Sensitivity	0.86	0.88	0.68	0.16	0.50
Specificity	0.97	0.92	0.88	0.83	0.56

とモデル化する．ここで $\mu_i = \mathrm{E}[X_i/(X_i+Y_i)]$ であり，ϕ は over-dispersion を表すパラメーターとなる．実際に X_i の分散は $\mathrm{Var}(X_i) = n_i\mu_i(1-\mu_i)\{1+(n_i-1)\phi\}$ と表される．ただし ϕ はグローバルパラメーターでは無く n_i の関数であり，3 点における fitting 結果から線形補間によって求めている様だ．コピー数 cn=1~3 と μ_i とは，ロジステックモデル

$$\mathrm{logit}(\mu_i) = \alpha + \beta_i + \gamma \mathrm{GC} \tag{5}$$

を仮定する．α はサンプル毎に決めるパラメーターであり，β_i は仮定するコピー数で 0.5 (loss)，1.0 (normal)，1.5 (gain) とし，GC は exon i の GC contents である．式 (2)-(5) から 3 つの状態 cn=1~3 を仮定して尤度を計算し，hidden Markov model を使って Viterbi アルゴリズムにより CNV を call する．

評価は immunodeficiency 患者 24 人 (Agilent 38Mb:15 人 50Mb:9 人) の末梢血から sequencing した WES データ (94bp の paired-end，novoalign で hg19 にマップ，平均 depth 37~81) と，1000 人ゲノム [9] から 12 人 (YRI:4，CEU:8) 分の WES データをダウンロードし，exomeCopy[25] 及び ExomeCNV[32] (CNA 検出用ツール) と比較している． ExomeDepth の予測結果は他よりも 20% 程度以上 DGV に登録されており (positive predictive value が高い)，また Conrad *et al.*[10] で aCGH から評価されている CNV の 75.2%を検出し，他の 52.8%，41.2%よりも sensitivity が高いと報告している．

臨床検体 (immunodeficiency 患者 24 人) から疾患と関連している候補 CNV として GATA2 と DOCK8 に 1 検体ずつ rare deletion を検出してカスタムチップや Sanger sequencing で確認している．

2.4 Coin *et al.* (2012)[7] (ExoCNVTest)

Case-control study データにおいて疾患と関連する common CNV を特定し，このコピー数を出力する．2 種類の主成分分析を利用してバイアスの低減を試みている．

あるサンプル k $(1 \le k \le N)$ において平均 depth で割られたある領域 i の j 番目の塩基 $(1 \le j \le M_i)$ の depth を r_{ijk} とする．この r_{ijk} を i 毎に主成分分析を行い (local PCA)，その第一主成分を $FPC_{i,k}$ とする．次に $FPC_{i,k}$ の主成分分析を更に行い (global PCA)，これを GPC_h $(h \ge 1)$ とする．この時 global PCA の上位はバイアスであると想定されるので，この寄与を除いた

$$FPC_i^{(H)} = FPC_i - \sum_{h=1}^{H} \frac{\langle FPC_i, GPC_h \rangle}{\langle GPC_h, GPC_h \rangle} GPC_h \tag{6}$$

を使ってロジスティックモデルにより p 値を計算する．ここで $\langle , \rangle$ は k に対するベクトルの内積を表している．コピー数の予測ではサンプル k の領域 i を 100 bp 毎の window j に分け，これを ra_{ijk} として式 (6) と同じ H の寄与を除いた

$$ra_{ij}^{(H)} = ra_{ij} - \sum_{h=1}^{H} \frac{\langle ra_{ij}, GPC_h \rangle}{\langle GPC_h, GPC_h \rangle} GPC_h \tag{7}$$

に対して cnvHap[6] を適用することにより予測する．

乾癬の 700 人対 800 人の case-control study において，NimbleGne 2.1M で read 長 90 bp の single read データ (depth 15x) を SOAPaligner[20] で NCBI build 36.3 にマップした．このデータに対して ExoCNVTest を適用したところ，$H = 0$ では χ^2 の median を期待値で割った inflation factor λ は 8.5 と非常に大きく，$H = 40$ とすると $\lambda = 0.92$ とすることが出来た．この時既知の deletion である *LCE3B* は $p = 7.2 \times 10^{-5}$ でランキングは 25 位，*LCE3C* は $p = 1.1 \times 10^{-4}$ でランキングは 33 位であった．$H = 40$ でのコピー数予測では accuracy は 97.4%であるが，missing 割合は 14.7%であった．このコピー数を使って関連解析を行うと $p = 5 \times 10^{-6}$ となった．

2.5 Li *et al.* (2012)[19] (CONTRA)

規格化された depth of coverage と control サンプル (multiple or matched) の baseline から，log-ratio を使って loss と gain を予測する．

サンプル s のライブラリーサイズを $L_s = N_s \times$read length$\times$percentage on target と定義する．ここで N_s

はマッピングされた全 read 数である．続いて塩基 b の adjusted coverage を raw coverage c_b から，コントロール集団の L_s の幾何平均 $\bar{L}$ を使って $d_b = c_b \times \bar{L}/L_s$ と定義する．更にコントロール集団における 10%トリム平均を $\bar{d}_s$ とする．塩基毎の log-ratio は BEDTools を使って閾値以上の depth (default 10 bp) を対象に計算し，領域毎の log-ratio (RLR) はこの平均値とする．RLR はライブラリーサイズ L_s が異ると log-coverage と線形の関係があるので，直線で fitting してこの bias を補正する．続いて同程度の log-coverage 区画において RLR$\sim N(\mu_d, \sigma_d)$ とモデル化する．μ_d は RLRs の平均とし，σ_d は log-coverage 区画のそれぞれで empirical SD を求め，これを直線補間することにより，d の関数として求める．$N(\mu_d, \sigma_d)$ から両側検定の p 値を計算し，FDR[4] で評価する．複数のターゲット領域を跨る CNV の検出は circular binary segmentation [27] (CBS) アルゴリズムを用いる．

シミュレーションデータ (Agilent SureSelect All Exon v2 の chr20 のデータから作成した depth 50x) を使って ExomeCNV[32](CNA 検出用ツール) と比較し，specificity は両方とも非常に高いが sensitivity は CONTRA の方が圧倒的に高かった (50-200 bp の CNV の sensitivity は CONTRA: 68%，ExomeCNV: 25%)．実データへの適用として 1000 人ゲノムから CEU の男性 5 人のデータ (NimbleGene V2) をダウンロードし，control をこの 5 人に 1 人を加えた計 6 人として CONTRA を実行した．これらのサンプルの CNV は HapMap[8] で評価されており，これを正解データとして評価したところ，$p = 0.01$ の設定で sensitivity: 86.8%，specificity: 95.4% であった．

2.6 Krumm *et al.* (2012)[15] (CoNIFER)

Singular value decomposition (SVD) を使った規格化により頻度 1%未満の稀な CNV を検出し，また既知の copy number polymorphic (CNP) の遺伝子型 call も行う．尚，稀な CNV の検出では exon target 領域が 3 つ以上となる CNV を call する．

CoNIFER のパイプラインでは，read を 36 bp 毎に分けて mrsFAST[12] でマッピングし，RPKM[26] を計算して median が 1 未満の領域を削除してから z-score に変換 (ZRPKM) する．この ZRPKM 行列に SVD を実行し，稀な CNV の検出では上位 12~15 位，CNP の call では上位 6 位までの主成分を除いて (特異値を 0 にして再構築)，SVD-ZRPKM 行列を作成する．続いて SVD-ZRPKM 行列において SD が 0.5 を越えたサンプル除き，再度 ZRPKM 行列を計算し，実際の call に用いる SVD-ZRPKM 行列を構築する．稀な CNV の call では SVD-ZRPKM において閾値-1.5 と 1.5 以上で連続する 3 箇所以上の条件を採用し，CNP では入力 (既知) の CNV 領域中で SVD-ZRPKM の平均値を使って集団中の頻度情報 (クラスタリングして最大頻度を基準にする) や対応する HapMap の WGS データを使ってコピー数を call する．Multi-allelic CNP の多くは segmental duplication 領域内にあり，40 コピーを越える多型もある．この様な多型の call ではコピー数を少なく call する．

実データへの適用では，aCGH や whole-genome shotgun sequencing，targeted clone sequencing，qPCR 等で既に CNV が調べられている HapMap[8] の 8 人の exome データ (NimbleGen v1，PE，90x) と，創始者 109 人分の Illumina Omni-1 データがある autism spectrum disorder (ASD) の 122 トリオの exome データ (NimbleGen v2，PE，76x) の解析を行った．ただし SVD のために control として NHLBI から 533 人の exome データ (NimbleGen v2，PE，81x) をダウンロードし，HapMap[8] の 8 人の解析では 533 人を加えて 12 主成分を除き，ASD の解析では 366 人を加えて 15 主成分を除いた．HapMap 5 サンプルの call 結果からは，positive predictive value (precision) は rare CNV で 86% (6/7)，CNP は 64% (16/25) であった．また Conrad *et al.*[10] で検出された CNV については，rare CNV は 5/5 で全て検出し，WGS データから予測された CNV とは 222/378 (59%) 領域で $r^2 \geq 0.9$ となった．ASD 122 トリオからの CoNIFER による rare CNV call では 117/124 (94%: positive predictive value) が Omni-1 や qPCR で確認され，sensitivity の評価は 83/109 (76%) であった．ExomeCNV[32](CNA 検出用ツール) と 4 人の HapMap データで比較したところ，ExomeCNV[32] は 63/450 (14%) しか既知の CNV と一致しなかったが，CoNIFER は 21/24 (88%) 一致した．

尚，492 人の NHLBI データを用いてゲノムへのマッピングで multiple mapping を許容する mrsFAST[12] と unique mapping モードの BWA[18] を比較したところ，BWA[18] の方が揺らぎが小さくなり 6 つの主成分を除けば良かったが，signal-to-noise ratio (SNR) は悪くなり予測精度が落ちたと報告している．

2.7 Klambauer *et al.* (2012)[13] (cn.MOPS)

本来 WGS データ用に開発された multi-sample 用ツール (cn.mpos) であるが，WES の CNV 予測ツールを比較した Yan *et al.* (2013)[38] で評価対象になっており，また文献中でも最後に適用を試みていると述べられ，実際にソフトウエアのマニュアルには WES データへの適用方法が記載されているため (exomecn.mpos) 調査対象に含めた．WGS，WES データ共に 5∼6 サンプル以上を同時に解析することが望ましいとしており，CNA の場合 (referencecn.mops) には matched pair を入力するとマニュアルに記載されている．

NGS データを Bowtie[16] でゲノムにマッピングし，サンプル毎にマッピングされた総 read 数で規格化する．規格化されたセグメント毎 (全ゲノム解析では 25 kbp，染色体 1 番のみの解析では 500 bp を使用．ただしソフトウエアマニュアルには，default は平均 100 カウント程度になる様にセグメント長を自動で探索すると記載されている．また WES の場合は exon や bait 位置を bed で入力.) の read カウントをコピー数毎に Poisson 分布 $P(\lambda)$ の和

$$p(x|\vec{\alpha},\lambda) = \sum_{i=0}^{n} \alpha_i P(x; \frac{i}{2}\lambda) \tag{8}$$

でモデル化する．ここで α_i はコピー数 i を含むサンプルの割合であり，λ はコピー数 2 の平均 read カウント数を表す．コピー数が 0 の場合はシーケンサーのエラー割合を反映したパラメーター $i = \varepsilon$ とする (文献中は 0.05)．式 (8) の N サンプルデータ x_k $(1 \le k \le N)$ に対する尤度の最大化では，α_i と λ を独立の確率変数としてその事後確率を

$$p(\vec{\alpha},\lambda|x) = \frac{p(x|\vec{\alpha},\lambda)p(\vec{\alpha})p(\lambda)}{\int p(x|\vec{\alpha},\lambda)p(\vec{\alpha})p(\lambda)d\vec{\alpha}d\lambda} \tag{9}$$

とモデル化する.$p(\vec{\alpha})$ の事前分布は Dirichlet 分布

$$p(\vec{\alpha}) = \frac{1}{B(\gamma)}\prod_{i=0}^{n}\alpha_i^{\gamma_i-1} \tag{10}$$

を仮定し，$i \neq 2$ に対して $\gamma_2 \gg \gamma_i \ge 1$ とすることで事前確率をコピー数 2 で極大化させておく．また λ の事前分布は一様分布とする．式 (8) を expectation maximization (EM) アルゴリズムで最大化することにより $\vec{\alpha}$ 及び λ を求める．ただし実際の推定ではコピー数 2 のサンプルが含まれる確率が 0 にならない様に $\gamma_2 = 1 + G$ として G を推定している．CNV の call では informative/non-informative (I/NI) 及び signed individual informative/non-informative (sI/NI) を

$$\begin{aligned}
\mathrm{I/NI}(\alpha) &= \sum_{i=0}^{n}\alpha_i|\log\frac{i}{2}| = \frac{1}{N}\sum_{k=1}^{N}\sum_{i=0}^{n}\alpha_{ik}|\log\frac{i}{2}|, \\
&= \frac{1}{N}\sum_{k=1}^{N}\mathrm{I/NI}(\vec{\alpha}_k), \qquad (11) \\
\mathrm{sI/NI}(\vec{\alpha}_k) &= \sum_{i=0}^{n}\alpha_{ik}\log\frac{i}{2}, \\
&\simeq \mathrm{sgn}\left(\sum_{i=0}^{n}\alpha_{ik}\log\frac{i}{2}\right)\mathrm{I/NI}(\vec{\alpha}_k), \qquad (12)
\end{aligned}$$

と定義し，独自に開発した fastseg アルゴリズムか CBS[27] でセグメント化する．ここで $\alpha_{ik} = \alpha_i P(x_k; i\lambda/2)/p(x_k|\vec{\alpha},\lambda)$, $\alpha_i = (\sum_k^N \alpha_{ik})/N$ である．セグメント化された領域においてサンプル間で call に違い (全サンプルで loss や gain となる場合は call しない) があり，且つ $\mathrm{sI/NI}(\vec{\alpha}_k)$ の median が 0.6 以上の場合に gain，-1 を以下で loss と call する．整数のコピー数の推定では，事前確率を α_i，尤度関数を $p(x|i) = P(x; \frac{i}{2}\lambda)$ として事後確率 $p(i|x)$ が最大となるコピー数とする．

既に CNV が解析されている HapMap サンプルの WGS データを使って評価したところ，既存のツール (MOFDOC[2]，EWT[39]，JointSLM[22]，CNV-seq[37]，FREEC[5]) よりも loss，gain 共に sensitivity (recall)，positive predicitive value (precision) が高かった．

cn.MOPS は R 上のソフトウエアとして実装されており，cn.mpos と exomecn.mpos では default パラメーターが異り，アルゴリズムに違いがある様ではなかった．

2.8 Fromer *et al.* (2012)[11] (XHMM)

疾患と関連している CNV は稀な CNV であることが想定されるため，頻度 5%未満となる様な rare な CNV の検出に最適化した WES データからの CNV 予測ツールとして XHMM は開発されている．想定しているデータは coverage 60x∼100x 程度の 50 サンプル以上の WES データセットである．

WES データからの CNV 予測パイプラインとして，WES 特有の様々な揺らぎやバイアスに対応するため，

1. BWA[18] でゲノムにマップし，local realignment，PCR duplicate の marking，base-quality-score recalibration を GATK[21] 等で行い，mapping quality 20 以上の depth をカウントして target 領域毎に平均 depth を計算する．

2. GC 含有量が 0.1 未満と 0.9 超の target 領域，RepeatMasker でマスクされる領域が 10%以上の

target 領域，10bp 未満及び 10kb 超の target 領域，coverage が全てのサンプルで 10x 未満及び 500x 超の target 領域を解析から除外.

3. サンプルにおける全 target 領域の平均 coverage が 50x 未満若しくは 200x 超，SD が 120 超となるサンプルを除外.

4. サンプルと target 領域毎の平均 depth による read-depth 行列 R を主成分分析し，$0.7/n$ 以上の分散となる K 成分の寄与を $R^* = R - \sum_{i=1}^{K} c_i c_i^T R$ と除く．ここで n はサンプル数であり，c_i は i 番目の主成分である．R^* を規格化された read-depth 行列と呼ぶ.

5. R^* において，SD が 30 超となる target 領域を除き，サンプル毎に Z スコアに変換する.

6. HMM により，diploid，deletion，duplication のセグメントを推定する．Exome では taget 領域間の距離に大きな差があるため，遷移行列には $f = e^{-d/D}$ の重みを導入する．ここで d は taget 領域間の塩基数であり，D は CNV 間の塩基数の期待値である．出力確率は分散 1 で平均が $-M$，0，M の正規分布を用い，Viterbi アルゴリズムで CNV を推定する.

7. 遺伝子型 call (実際には deletion，diploid，duplication のみを call し，これらの組み合わせ出は無い) は forward-backward アルゴリズムから 7 つ状態 (exact deletion，some deletion，no deletion，left deletion breakpoint，right deletion breakpoint，not diploid，diploid) の事後確率を計算し，これらを組み合わせて deletion を call する (duplication も同様)．遺伝子型 call の特徴として diploid (2 copy) も call 対象としており，そのため no call も出力する.

としている.

パイプライン評価方法としては，進行中の統合失調症研究プロジェクトから，ブルガリアの 90 人のトリオ (30 家系の両親とその子) とスウェーデンの case-control study サンプル 1,017 人分の末梢血 DNA からの WES データと Affymetrix SNP チップデータを用いた．90 人のトリオデータは Mendelian transmission rate が 50%となり且つ *de novo* CNV が殆んど無いと言う条件で HMM のパラメーター推定に使い，このパラメーターで 1,017 人の CNV call を行った．頻度 1%未満となる 2,315 CNV (80% 以上は 100kb 未満の CNV) を検出し，Affymetrix SNP チップデータから call した頻度 1%未満の結果の内 exon target 領域が 1 つ以上含まれている 544 CNV と比較したところ，367 CNV (67%) を検出していた.

主成分分析を用いた規格化を採用している同様のソフトウエアである CoNIFER [15] と (XHMM のパラメーター fitting で用いた)90 人トリオのデータで比較したところ，CoNIFER[15] の方が rare CNV call 数及び *de novo* CNV が多く，Mendelian transmission rate も 29%と低かった.

2.9 Magi *et al.* (2013)[23] (EXCAVATOR)

集団の control から CNV を予測する pooling モードと matched pair を使って CNA を予測する somatic モードあり，3 ステップによる規格化の後に 5 つの状態 (0∼3, 4 コピー以上) として予測する.

予測に使う指標は，exon mean read count (EMRC) であり，$EMCR_i = RC_i/L_i$ で定義される．ここで RC_i は exon i にマップされた read 数であり，L_i は exon i の長さである．この $EMCR_i$ に対して GC 含有量，mappability[14]，exon 長らの bin 毎に中央値で割ることにより規格化し，更に control の値 (pool 若しくは matched control) で割って $\log_2$ を取った値 ($\log_2$ ratios) を指標として使用する．この $\log_2$ ratios を彼らが開発した heterogeneous shifting level model (HSLM) を使って segment 化し，FastCall algorithm[3] で 5 つの状態 (0∼3，4 コピー以上) に call する．これらのアルゴリズムの評価では，Affymetrix SNP 6.0 チップを使った 1000 人ゲノム [9] の 8 人 (CEPH 1 人，Yoruba 7 人) のデータを使って調べている.

集団データは 1000 人ゲノム [9] の 20 人 (CEU:7，JPT:7，YRI:6) BAM ファイル (SureSelect All Exon V2, 83x) をダウンロードして評価した．Control を 1 人の YRI (107x) として染色体 1 番と 4 番のみを使って XHMM[11]，CoNIFER[15]，ExomeCNV[32](CNA 検出用ツール) と比較した．ただしバイアスの除去に PCA を使っている XHMM[11]，CoNIFER[15] では 80 サンプル分のデータを加えて解析した．評価方法としては，同じサンプルに対して過去に解析した McCarroll *et al.*[31] (96 common，4 rare) 及び Conrad *et al.*[10] (116 common，4 rare) で検出された CNV に対して recall (sensitivity) と precision (positive predictive value) を全体，common，rare で評価した (図 1)．この結果から Conrad *et al.*[10] データにおいて EXCAVATOR が他よりも良い成績であることから，他のツールを outperform して

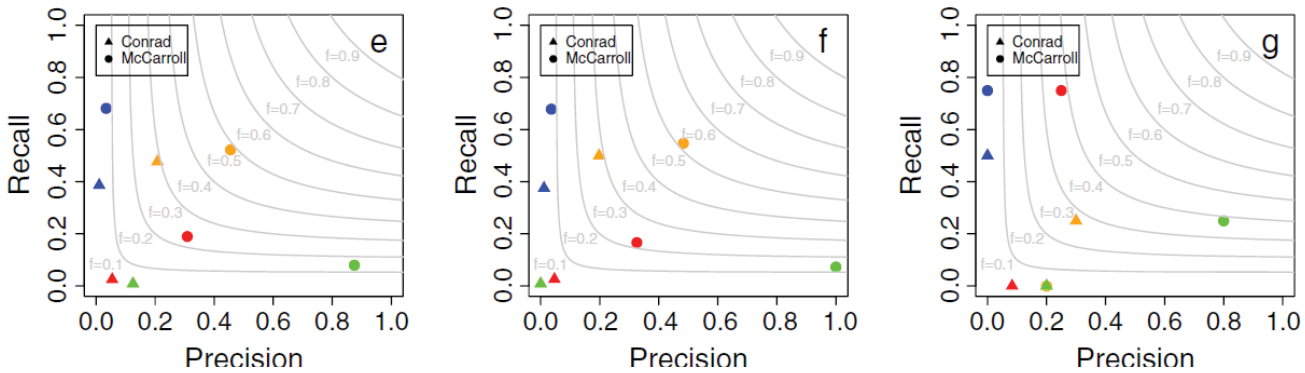

図 1 Magi *et al.*[23] の figure 3 の (e), (f), (g). 左図:全体, 中央図:common, 右図:rare. 橙:EXCAVATOR, 赤:XHMM, 緑:CoNIFER, 青:ExomeCNV.

いると主張している.

続いてメラノーマのデータを使って評価してるが, CNA は評価対象外のため割愛する (この解析では control を 6 サンプル分プールしている).

知的障害者のきょうだい 2 サンプル (Name1,Name2) の WES (Trueseq, 63x, PCR duplicates を削除, base quality recalibration, local realignment 済) を実施し, control としてヨーロッパ人の健常者 1 サンプルの WES データ (SRA040093) をダウンロードして Name1,Name2 の CNV を推定した. それぞれ 29 CNV, 24 CNV を検出し, DGV にはそれぞれ 22, 17 CNV が登録されていた.2q11.1-2q11.2 の deletion が 2 検体で共に検出され, Affymetrix SNP 6.0 チップの結果とも一致した.

2.10 Yan *et al.* (2013)[38] (方法論の比較)

CoNIFER[15], cn.MOPS[13](WGS 用), exomeCopy[25] 及び ExomeDepth[29] を対象として, がんの cell サンプルに対して aCGH データと WES データ (75 base paired end でサンプル当り 73~183M read, capture 効率 39~62%) からの予測結果を比較している.

乳がん 16 サンプルの cell line に対して, Agilent SurePrint G3 Human CGH Microarray (963,029 プローブ) による aCGH と Illumina TrueSeq による WES を行った.

aCGH データは GeneSpring の Aberration Detection Method 2 (ADM2) で CNV を予測した. WES データは BWA[18] で hg19 にマップし, GATK[21] で local realignment 及び base quality score recalibration を行い, mapping quality が 20 未満の read を削除し, base quality が 20 未満の塩基を call 対象から外した (PCR duplicates を削除した記述は無し).

全 16 サンプルから aCGH により 5,225 CNV が予測され, 一方 WES データからは CoNIFER[15] 267, cn.MOPS[13] 1,214, exomeCopy[25] 3,398, ExomeDepth[29] 1,581 CNV がそれぞれ予測された. aCGH の結果を gold standard として感度 (sensitivity, true positive rate) を評価しているとあるが, 精度 (precision, positive prediction value) を評価している様にに思える. また aCGH では deletion 数と duplication 数の違いは統計的に有意ではなかったが, WES データでは exomeCopy[25] 以外は duplication を有意に多く予測する結果が得られている. 結論として, false positive rate が低い CoNIFER[15] と WGS データ用の cn.MOPS[13] の使用を勧めている.

筆者の感想

乳がん 16 サンプルの cell line には germline の CNV よりも somatic な CNA が長大 (染色体単位や短腕, 長腕単位等) に多数含まれているはずであり, また ploidy も 2 であるとは限らず, CNV 予測ツールの評価には著しく不適当である. また PCR duplicates を除かずに duplication bias を評価するのも不適当であろう. 更に sensitivity に触れずに precision のみでツールを評価することは, この評価結果が参考になるケースを大きく制限している.

2.11 Renjie *et al.* (2014)[30] (方法論の比較)

XHMM[11], CoNIFER[15], ExomeDepth[29], 及び CONTRA[19] を対象として, WGS データからの予測結果の再現性, tagSNP を用いた CNV データベース (common CNV) との比較, Mendelian エラー割合のチェック, deletion 中の hetero single-nucleotide variant (SNV) の有無のチェックを行っている.

測定に用いた DNA が末梢血由来なのか, cell line 由来なのか記載されていない. 9 trio (27 人) を含む 33 人分の exome (Agilent SureSelect 50Mb) データ (平均 73x) と, この内の 13 人分 (非家系) の WGS デー

表 2 4 つのツールの評価結果. 数字は%.

評価方法	XHMM[11]	CoNIFER[15]	ExomeDepth[29]	CONTRA[19]
WGS 予測の再現割合 (高い方が良い)	7.67	3.18	18.17	4.32
Common CNV の検出割合 (高い方が良い)	13.58	15.64	34.57	15.64
Mendelian error rate (低い方が良い)	20.00	10.68	16.11	56.25
Heterozygosity check (1kb 超，低い方が良い)	64.10	18.75	34.02	40.00

表 3 WES データからの CNV 予測ツール一覧

ツール名	補正	Over-dispersion 対策	Call 方法
exomeCopy[25]	GC，depth，window 長	負の 2 項分布	HMM
ExomeDepth[29]	GC	beta-binomial	HMM
ExoCNVTest[7]	Local PCA	Global PCA	cnvHap[6]
CONTRA[19]	塩基数	Depth に依存する正規分布	CBS[27]
CoNIFER[15]	外れ値の削除，SVD		連続する 3ヶ所以上が閾値越え
XHMM[11]	外れ値の削除，対象領域の限定，PCA (SVD)		HMM
EXCAVATOR[23]	GC，mappability，exon 長	HSLM	FastCall[3]

タ (~38x) を取得した．これらのデータを BWA[18] で hg19 にマップし，PCR duplicates を Picard で除いた．WGS データについては ERDS[40] と CNVnator[1] で CNV 予測を行い，exon の 50%以上を被覆する CNV 領域を比較対象として exon 単位の和集合を正解データとした (median 2,802 exon).

4 つの予測ツールの call を比較すると，CoNIFER[15] は最も CNV 予測数が少なく (中央値 13)，CONTRA[19] は最も予測数が多い (平均 811) が 96%は 1kb より短い CNV だった．この 4 つのツールの 4 つの方法による評価結果を表 2 に纏めた．Common CNV の検出割合では，HapMap[8] の CEPH 集団で得られている CNV と $r^2 \geq 0.8$ となる tagSNPs を hg19 上にて 31,104 ペア抽出し，この tagSNPs が call されている時の CNV 検出割合を算出している．Mendelian error rate では，両親の約半分の CNV が子に遺伝することを利用してエラー割合を評価し，Heterozygosity check では 1kb 以上と予測された deletion 領域に hetero となる SNV がある場合の割合を求めている．

2.12 文献調査のまとめ

WES データ用に開発された CNV 予測ツールを表 3 にまとめた．Window 単位で予測する exomeCopy[25] 及び ExoCNVTest[7] を除いて他は全て exon 単位の read count や read depth による予測ツールであった．ただし ExomeDepth[29] は exon 内の indel も call している．WES データは WGS データに比べてバイアスや揺らぎが大きく，control として使用するデータセットには解析対象サンプルと極めて近い実験であることが求められ，加えて over-dispersion への対策も必要となる．ただし「比の統計量」は文献 [32] で指摘されている様に[2] 元々裾野が広く外れ値が多くなるが，これは read count を正規化 (総 read 数で割る) して予測を行う WGS からの予測ツールでも問題となる．

文献調査の結果から，本調査で WGS データ及び Omni2.5 チップデータとの比較を行うツールを表 4 にその理由と共に示した．

表 4 評価対象 WES データからの CNV 予測ツール

ツール名	理由
ExomeDepth[29]	Renjie *et al.* (2014)[30] で総合的に成績が良かった.
XHMM[11]	疾患との関連探索では rare CNV の予測精度が求められる.
EXCAVATOR[23]	発表が新しく，ツールの比較論文で評価されていない.

[2]例えば正規分布 $N(0,1)$ に従う独立な確率変数 X，Y の比 $R = X/Y$ は平均や分散が定義出来ない Cauchy 分布に従い，中心極限定理の対象外となる．Cauchy 分布からの標本平均は Cauchy 分布に従い，t 検定等では偽陽性を制御出来ない.

3　比較の方法

3.1　比較対象データ

2013 年 3 月に Illumina (Sun Diego) に外注した末梢血由来の 28 人分の whole genome sequencing (WGS) データ (ゲノムを 2,861,052,551 bp として平均 depth 44) とそれに付随する Omni2.5 チップデータから，

1. ERDS[40] による WGS データからの CNV 予測 (ERDS 予測データ)
2. CNVnator[1] による WGS データからの CNV 予測 (CNVnator 予測データ)
3. PennCNV[35] による SNPs チップからの CNV 予測 (Omni-CNV 予測データ)

を行った．以降では ERDS 予測データと CNVnator 予測データをまとめて「WGS-CNV 予測データ」と呼ぶことにする．この同じ 28 サンプルに対して 2013 年 4 月に WES を行ったデータ (Agilent SureSelect V4+UTR+lincRNA:ターゲット長 121,466,001 bp，平均 depth 76) を用いて，ExomeDepth[29]，XHMM[11]，EXCAVATOR[23] による予測 (WES-CNV 予測データ) を行った．ただし XHMM[11] では control 集団として本プロジェクトにて 2012 年に同じ bait で sequencing した 192 人分の WES データ (平均 depth 84.4) を用いた．これらのデータの depth の分布を図 2 に示した．尚，本調査では比較対象は常染色体のみとしている．

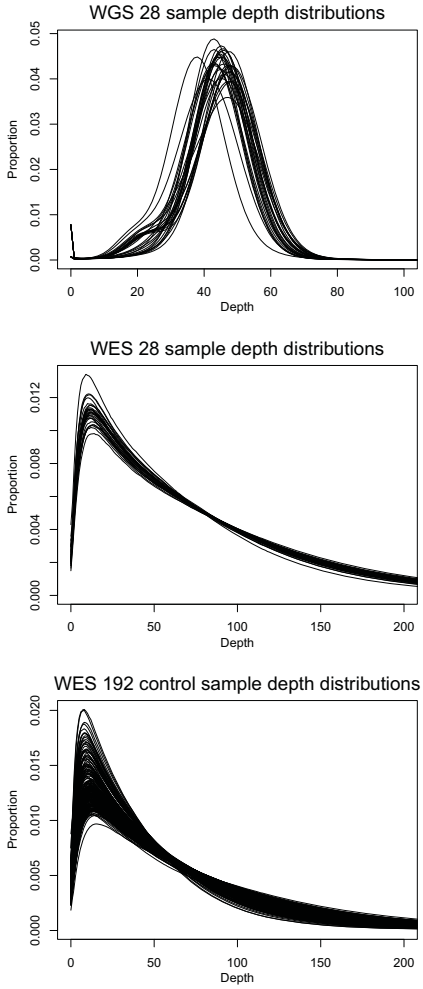

図 2　解析に用いた WGS 及び WES データの depth-of-coverage の分布．上図:WGS 28 サンプル，中央図:WES 28 サンプル，下図:control に用いた WES 192 サンプル．

3.2　WGS-CNV 予測データ

WGS データからの CNV 予測では，東北メディカル・メガバンク機構で行われた調査結果から ERDS[40] と CNVnator[1] を採用することにした．ただし集団データを同時に解析する方法は調査対象から外れていたため，本解析でも個人毎に予測を行った．

3.2.1　ERDS[40] の実行

ERDS[40] は GATK[21] の HaplotypeCaller でサンプル毎に call した結果 (SHC) を特に SNV 抽出を行わずにそのまま-v オプションで指定し，segmental duplication を指定する「–sd b37」オプション (UCSC の fasta Name にファイル名を変更) を指定して実行した．

3.2.2　CNVnator[1] の実行

CNVnator[1] はサンプル毎に bin_size を 100 で実行した．Mapping quality が 0 の read 由来の CNV call フラグ q0 のための「-unique」オプション[3]は指定しなかった．尚，全サンプルで最初の行に出力される chr1:1-10000 の loss は無視した．

3.2.3　予測結果

ERDS[40] 及び CNVnator[1] で call された loss 及び gain のセグメント数及びセグメント長の分布を図 3 に示した．またサンプル毎のセグメント数を図 4 に示した．ERDS[40] で call された最短長は loss が 109 bp，gain

[3]28 サンプル中 3 サンプルでエラーとなり実行出来なかった．Picard で BAM ファイルを修正しようとするとファイルサイズが非常に巨大になって終了しなくなったため，用意した BAM ファイルに異常があったと考えている．

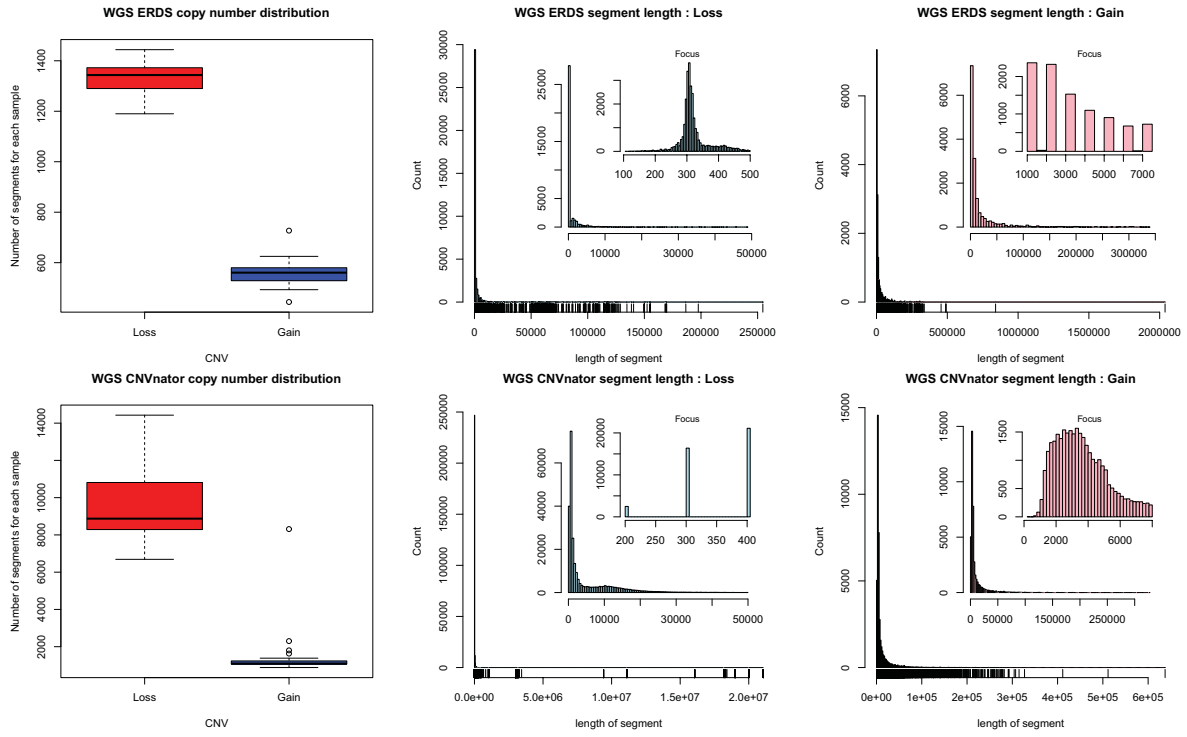

図 3 28 人の WGS データから CNV を予測した結果. 左図:サンプル毎の loss と gain のセグメント数の boxplot，中央図:全サンプルにおける loss の長さの分布，右図:全サンプルにおける gain の長さの分布. 上段:ERDS[40)]，下段:CNVnator[1)].

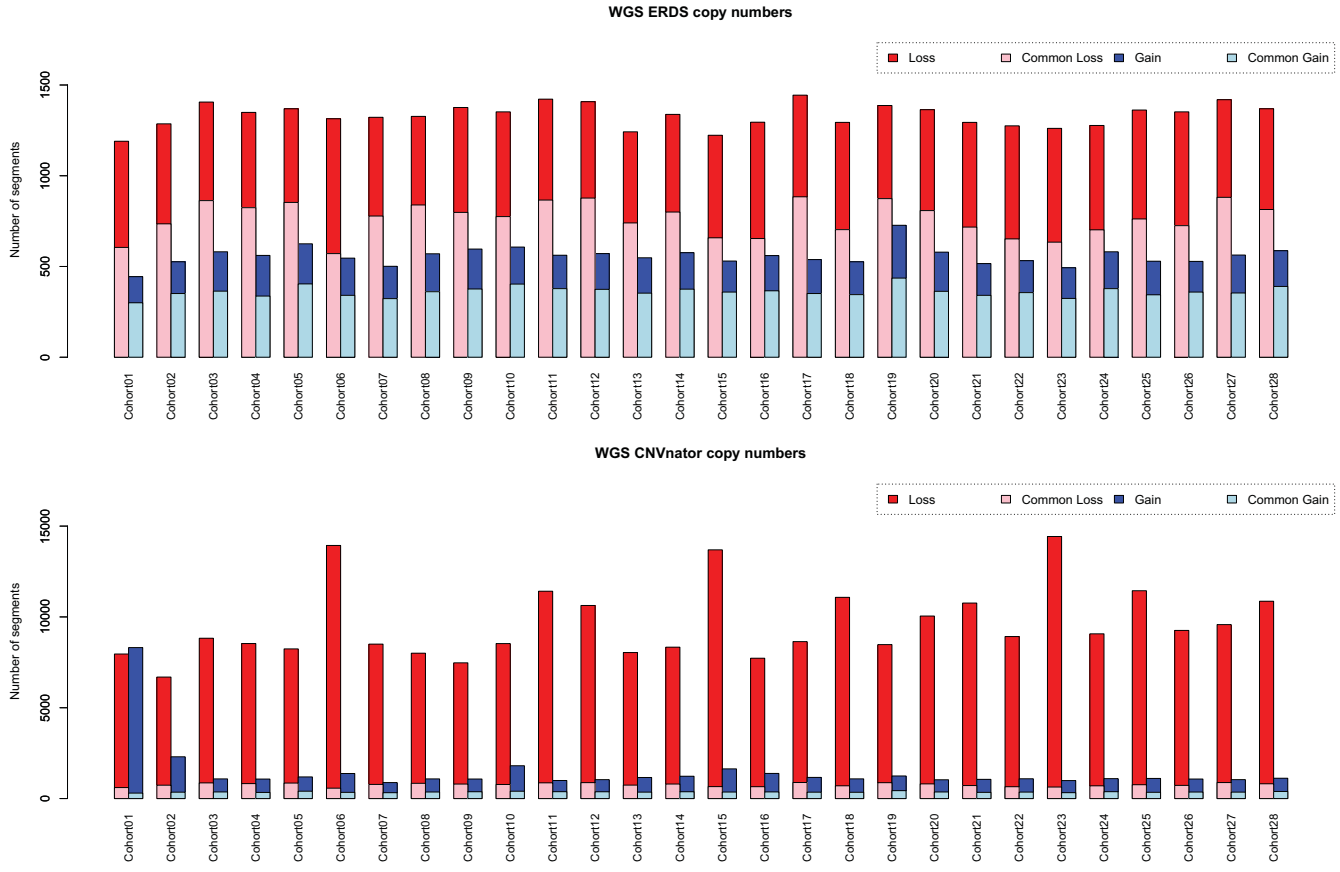

図 4 WGS データから call されたサンプル毎の CNV 数. 上段:ERDS[40)]，下段:CNVnator[1)]. 他方と一致した call を色分.

表 5 ERDS[40] で予測された長い CNV. Loss: 170 kbp 以上，gain: 400 kbp 以上.

CNV	chr	start	end	length (bp)	Sample Name
Loss	19	43,292,601	43,547,200	254,600	Cohort15
	19	43,589,401	43,776,000	186,600	Cohort28
	19	56,346,801	56,544,400	197,600	Cohort25
Gain	8	3,686,001	5,724,000	2,038,000	Cohort25
	10	42,597,001	43,053,000	456,000	Cohort14
	12	33,723,001	34,564,000	841,000	Cohort05
	21	10,698,001	11,188,000	490,000	Cohort09, Cohort15, Cohort25

表 6 CNVnator[1] で予測された長い CNV. Loss: 20 Mbp 以上，gain: 300 kbp 以上. 全サンプルで call されている最初の 1 番染色体の loss は centromere を跨いでいる.

	chr	start	end	length (bp)	Sample Name
Loss	1	121,485,401	142,535,400	21,050,000	全サンプル
	15	1	20,046,000	20,046,000	Cohort07
	15	1	20,038,300	20,038,300	Cohort15
	15	1	20,037,600	20,037,600	Cohort21
	15	1	20,022,700	20,022,700	Cohort11, Cohort26
	15	1	20,022,500	20,022,500	Cohort03, Cohort04, Cohort19
	15	1	20,000,200	20,000,200	Cohort16
	15	1	20,000,100	20,000,100	Cohort09, Cohort25
Gain	8	3,818,501	4,122,600	304,100	Cohort25
	8	4,479,401	4,890,800	411,400	Cohort25
	8	4,928,701	5,440,200	511,500	Cohort25
	12	33,924,201	34,561,700	637,500	Cohort05
	14	20,117,201	20,424,500	307,300	Cohort10
	16	33,306,001	33,633,000	327,000	Cohort20
	21	10,770,601	11,085,500	314,900	Cohort09

表 7 WGS 28 サンプルデータにおける ERDS[40] と CNVnator[1] の CNV call の重なり

		Only		Total		
CNV	Common	ERDS	CNVnator	ERDS	CNVnator	ERDS∪CNVnator
Loss	21,392	15,925	247,712	37,317	269,104	306,421
Gain	10,106	5,497	30,642	15,603	40,748	56,351

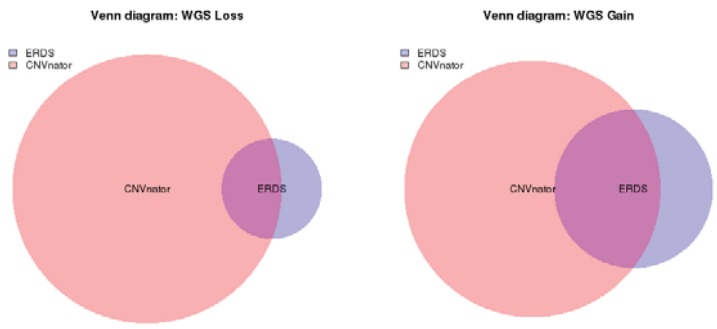

図 5 WGS-CNV 予測データのベン図. 左図:loss，右図:gain.

が 1,000 bp であり，CNVnator[1] では両方とも 200 bp であった (参考 Alu:約 300 bp，LINE-1:6.1 kbp). 一方で長い CNV と予測された上位を表 5，6 に示した.

WES からの予測結果らとの比較を容易にするために，CNV 予測結果は loss (deletion) と gain (duplication) のみの 2 値として取り扱った. ERDS[40] と CNVnator[1] の 28 サンプル分の call の重なりを表 7 及び図 5 に示した. ここで CNV 領域は loss 若しくは gain の判定が一致し，領域が 1bp 以上重なれば一致とした. ただし一つの CNV call 領域に複数個の他ツールの CNV call

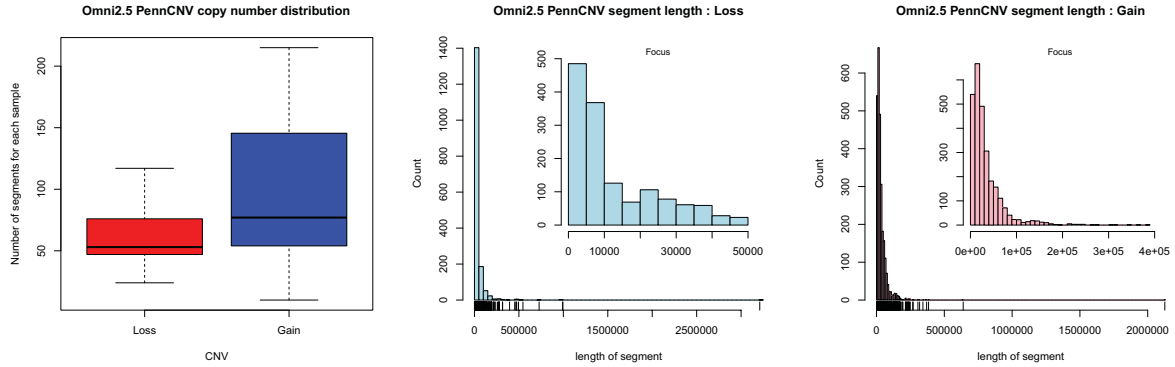

図 6 Omni2.5 チップデータから PennCNV[35] で予測した結果. 左図:サンプル毎の loss と gain のセグメント数の boxplot, 中央図:全サンプルにおける loss の長さの分布, 右図:全サンプルにおける gain の長さの分布.

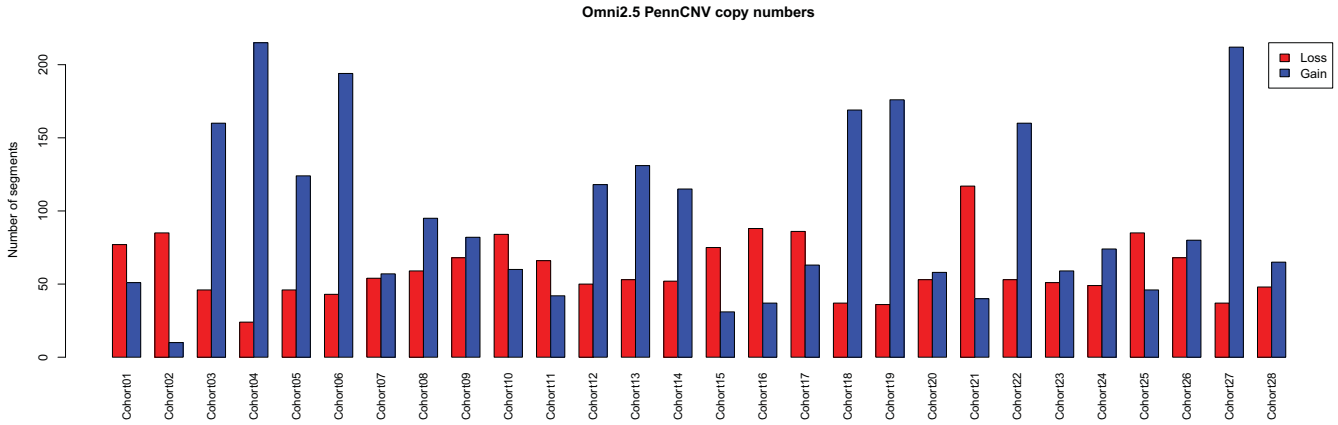

図 7 Omni2.5 チップデータから PennCNV[35] で予測したサンプル毎の loss と gain のセグメント数.

表 8 PennCNV[35] で 500kbp 以上と推定された CNV. n はコピー数. 19 番染色体の loss は centromere を跨いでいる.

chr	start	end	start SNP	end SNP	#SNPs	n	length (bp)	Sample Name
8	7,154,320	7,881,478	kgp4421102	rs7387061	14	1	727,159	Cohort03
8	7,336,103	7,881,478	kgp22800643	rs7387061	7	1	545,376	Cohort14
8	3,686,828	5,812,995	kgp4951559	kgp1993221	4,719	3	2,126,168	Cohort25
9	68,996,222	69,989,331	kgp2982617	kgp22790648	34	1	993,110	Cohort02, Cohort07
12	34,213,740	34,853,011	kgp19099743	rs12315121	324	3	639,272	Cohort05
19	24,594,797	27,804,863	kgp22760224	kgp22805080	23	1	3,210,067	Cohort15

領域が存在する場合や複数領域が互いにオーバーラップしながら続いた場合は, オーバーラップ領域が最も長い CNV call ペアを判別対象とし, 次にこのペアを除いて同様の手順を繰り返した.

3.3 Omni-CNV 予測データ

Illumina (San Diego) にて Omni2.5 チップで測定した B-allele frequency (BAF) 及び log R ratio (LRR) を用いて PennCNV[35] によって予測した CNV の loss 及び gain のセグメント数, 及びセグメント長の分布を図 6 に示した. またサンプル毎の loss 及び gain のセグメント数を図 7 に示し, CNV 長が 500kb 以上と予測された CNV のリストを表 8 に示した. これまでの経験では日本人集団に対する PennCNV[35] の出力結果は loss の方が gain よりも倍以上多いことが多いが, gain の方が多い結果となった. これは San Diego のデータのためサンプルクラスターではなく default クラスターを用いて BAF と LRR を計算していることが原因である可能性がある (loss はヘテロ SNP の存在で絞れるが gain は除けない). 尚 PennCNV[35] は default では 3 SNPs 以上続く領域のみを call するため, 以降の比較では call された CNV 領域に 3 SNPs 以上含まれている予測結果のみを対象とした.

3.4 WES データからの予測方法

Control 集団として本プロジェクトにて同じ bait (Agilent SureSelect V4+UTR+lincRNA:ターゲット長 121,466,001 bp) で sequencing した 192 人分の WES データ (平均 depth 84.4) を必要に応じて用いて，各ソフトウエアの default 値にて実行した.

3.4.1 ExomeDepth[29] の実行

Control サンプル 192 人と評価対象 28 サンプルを用いて,「aggregate reference」を評価しているサンプル以外の 219 人で作成して (1 人対 219 人) 推定を行った. また評価対象 exon 領域がソフトウエアに付属していたため，bait target 領域で無く提供されていた exon 領域 (185,130 exons，全長 32,091,604 bp) を使った. ExomeDepth は 2 塩基以上の indel も call するため，複数以上の exon を含む call と exon の半分以上を含む call のみを比較対象とした.

3.4.2 XHMM[11] の実行

Control サンプル 192 人と評価対象 28 サンプルの合計 220 人で実行し，28 人分の予測結果を xhmm_DATA.vcf ファイルから抽出した. 尚，filtering の段階で除かれたサンプルは無かった.

3.4.3 EXCAVATOR[23] の実行

評価対象 28 サンプルを 1 対他の 27 人として (192 サンプルの WES データは使用せず[4]) EXCAVATOR package v2.2 で実行した.

3.5 Omni-CNV 予測データと WGS-CNV 予測データの比較方法

WGS-CNV 予測データから，今回 PennCNV[35] による解析で用いた Omni2.5 チップデータのプローブが 3 つ以上含まれている領域を抽出した. ERDS 予測データ及び CNVnator 予測データのそれぞれにおいて比較対象となった loss 及び gain 領域数をサンプル毎に図 8 に示した. 一方で PennCNV[35] の予測結果は全て解析対象とした.

[4] 192 サンプル及び 192+27 サンプルを control とした時の方が PennCNV や ERDS，CNVnator と一致した CNV の数が loss, gain 共に減った.

3.6 Omni-CNV 予測データと WES-CNV 予測データの比較方法

Omni-CNV 予測データからは，bait target 領域が半分以上含まれている領域を抽出し，ExomeDepth[29]，XHMM[11]，及び EXCAVATOR[23] の予測結果と比較した. ExomeDepth[29] は独自の exon 情報を使っているが，評価対象領域は他の 2 つと同じにした. Omni-CNV 予測データにおいて比較対象となった loss 及び gain 領域数をサンプル毎に図 9 に示した.

精度評価としては，Omni-CNV 予測データを正解として，感度 (sensitivity; SN) と陽性的中率 (positive predictive value; PPV) を表 9 の様に評価することにした. ここで CNV call が一致するとは,

1. CNV loss 若しくは gain の判定が同じ
2. CNV 領域に 1 塩基以上重なりがある
3. 複数の CNV 領域が互いに重なる場合，最もオーバーラップが長いペアで評価する (最長が loss と gain で 2 つ以上ある場合は一致を優先). 続いてこのペアを除いて更に CNV 領域が互いに重なる場合は，逐次的に最もオーバーラップが長いペアを評価して行く.

とした.

3.7 WGS-CNV 予測データと WES-CNV 予測データの比較方法

WGS-CNV 予測データから，bait target 領域が半分以上含まれている領域を抽出した. ERDS 予測データ及び CNVnator 予測データのそれぞれにおいて比較対象となった loss 及び gain 領域数をサンプル毎に図 10 に示した. ExomeDepth[29] は独自の exon 情報を使っているが，評価対象領域は他と同じ「bait target 領域が半分以上」を使用した.

WGS-CNV 予測データにそれぞれ ExomeDepth，XHMM，EXCAVATOR の call 結果を加えたベン図に加え，前節と同様に WGS-CNV 予測データを正解とした評価表を表 9 の形式で作成した.

3.8 Rare CNV の抽出

ERDS 予測データ，CNVnator 予測データ，Omni-CNV 予測データのそれぞれにおいて，他のサンプルと全く重ならない CNV を rare CNV と定義した. 各予測データセットにおける rare CNV 数を表 10 に，サンプル毎

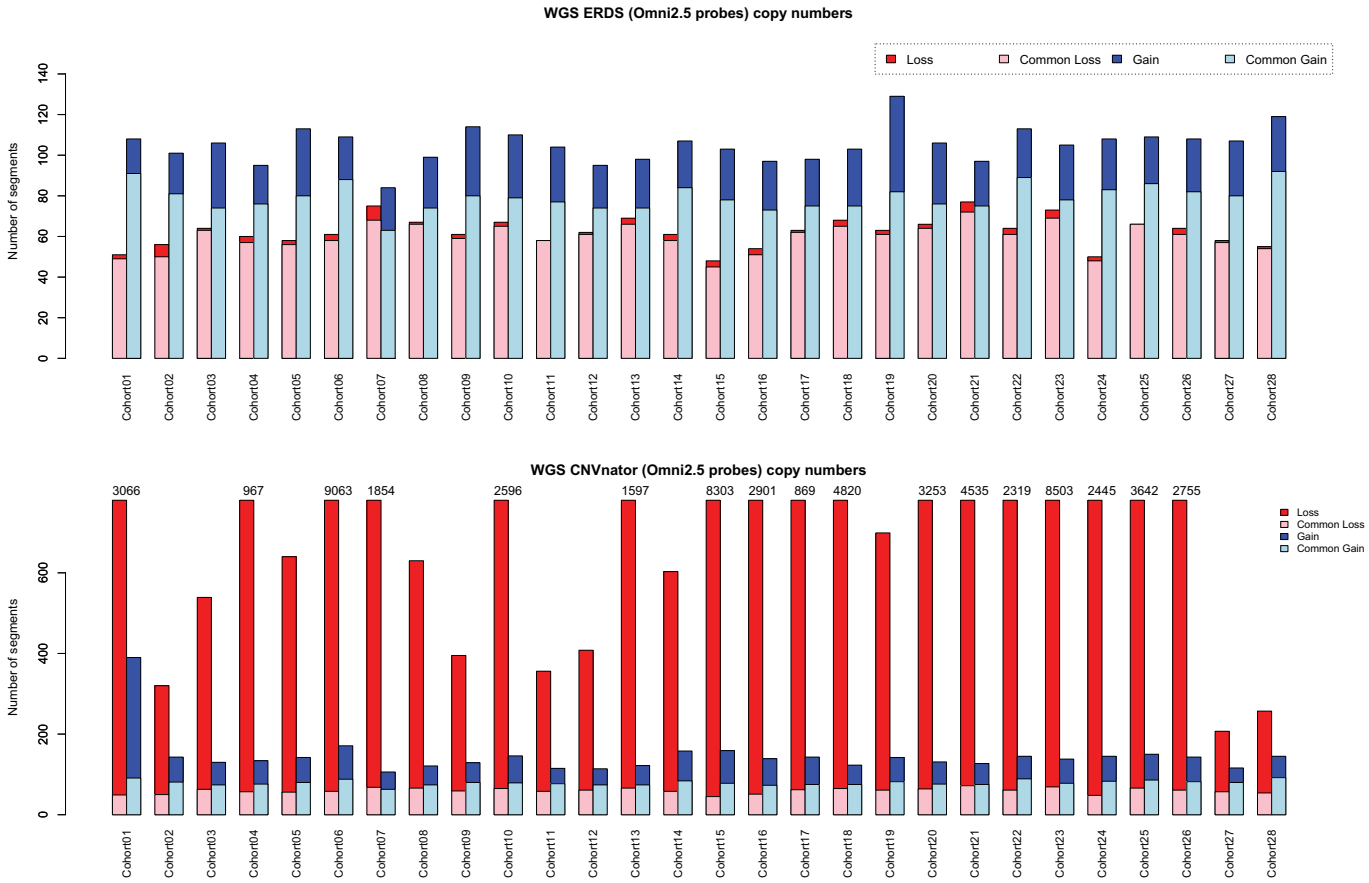

図 8　WGS-CNV 予測データの内，Omni CNV call と比較対象となった領域数．上段:ERDS[40]，下段:CNVnator[1]．CNVnator[1] は loss が gain に比べて非常に多いため，縦軸のスケールを越えた場合は数字で示した．

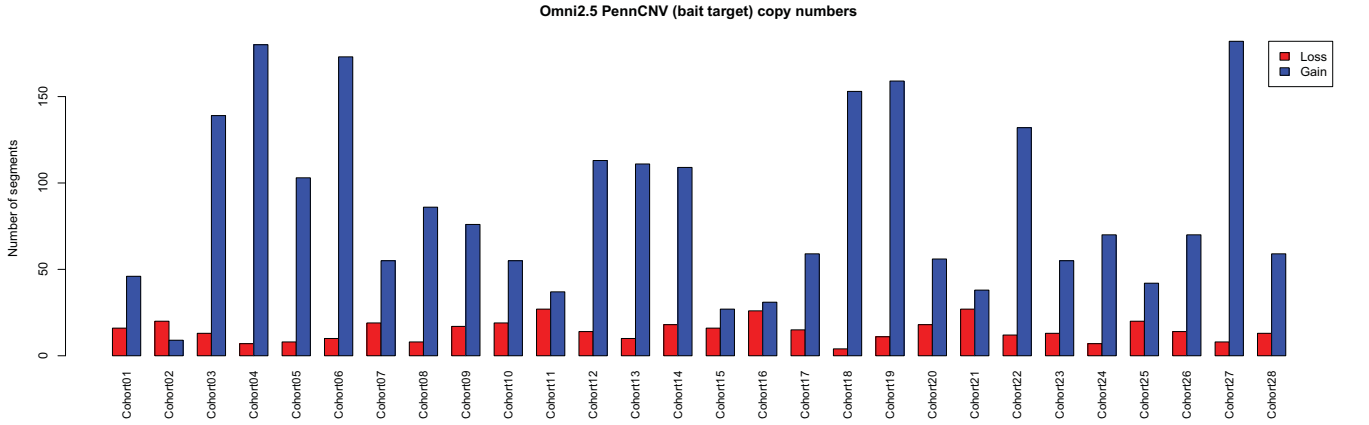

図 9　Omni-CNV 予測データの内，WES CNV call と比較対象となった領域数．

表 9　WGS 若しくは Omni チップのセグメント単位の推定結果を gold standard とした時の評価表．LR:loss region，NLR:not loss region，GR:gain region，NGR:not gain region．

		WGS/Omni loss call				WGS/Omni gain call		
	call	LR	NLR	Prop.	call	GR	GLR	Prop.
WES	LR	TP	FP	$PPV = \frac{TP}{TP+FP}$	GR	TP	FP	$PPV = \frac{TP}{TP+FP}$
	NLR	FN	-	-	NGR	FN	-	-
	Prop.	$SN = \frac{TP}{TP+FN}$	-			$SN = \frac{TP}{TP+FN}$	-	

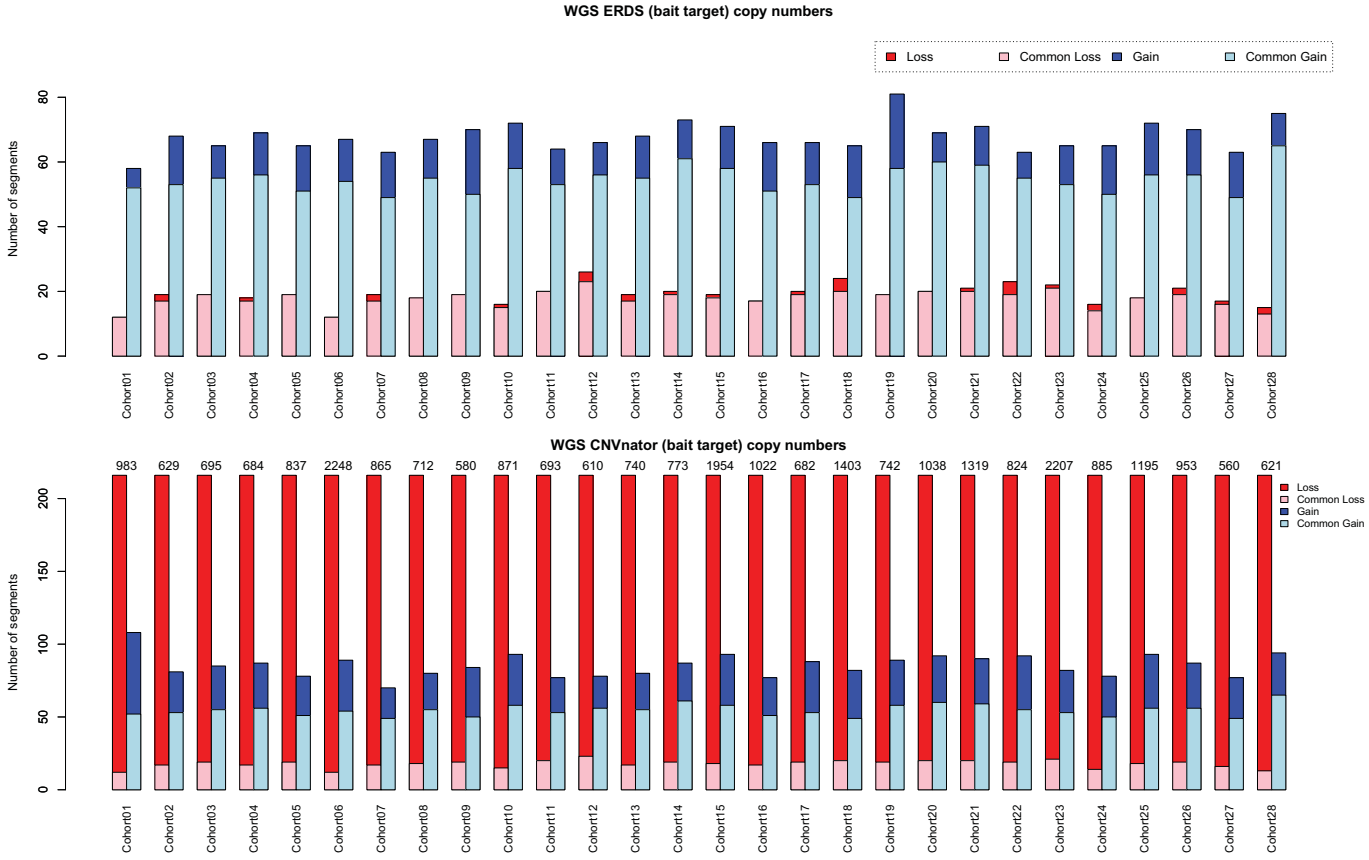

図 10　WGS-CNV 予測データの内，WES CNV call と比較対象となった領域数．上段:ERDS[40]，下段:CNVnator[1]．CNVnator[1] は loss が gain に比べて非常に多いためこれを数字で示した．

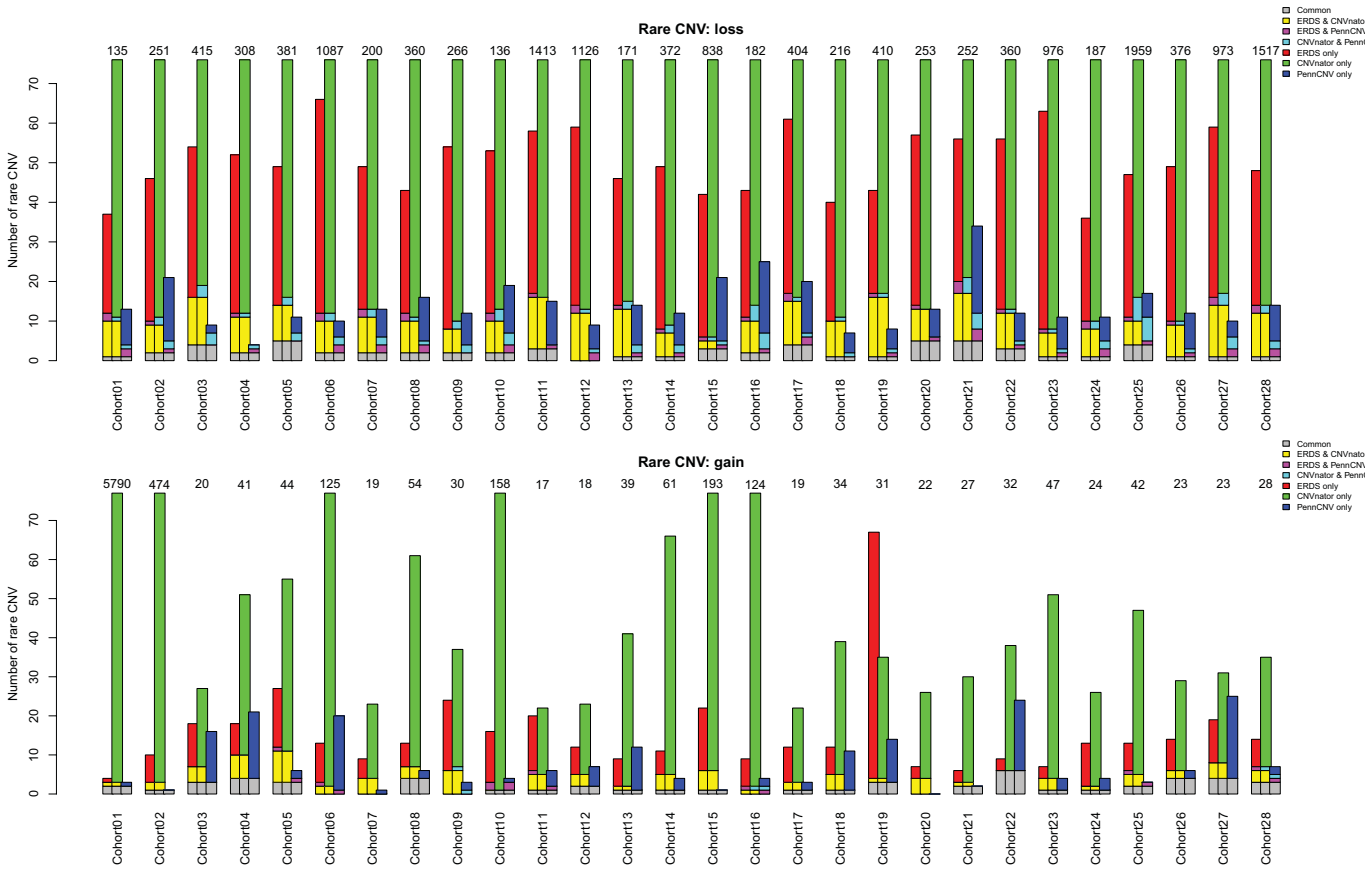

図 11　各予測ツールにおいて 1 サンプルのみで call された rare CNV 数をサンプル毎に，左から ERDS[40]，CNVnator[1]，PennCNV[35] の順に他のツールと一致した call を色分けして表示．上段:loss，下段:gain．上部の数字は CNVnator の数．

表 10　Omni-CNV 予測データと WGS-CNV 予測データの rare CNV call の重なり．ER:ERDS[40]，CN:CNVnator[1]，PC:PennCNV[35]．

					Only			Total		
CNV	ER∩CN∩PC	ER∩CN	ER∩PC	CN∩PC	ER	CN	PC	ER	CN	PC
Loss	62	253	36	52	1,064	15,524	243	1,415	15,891	393
Gain	52	82	8	3	286	7,559	155	428	7,696	218

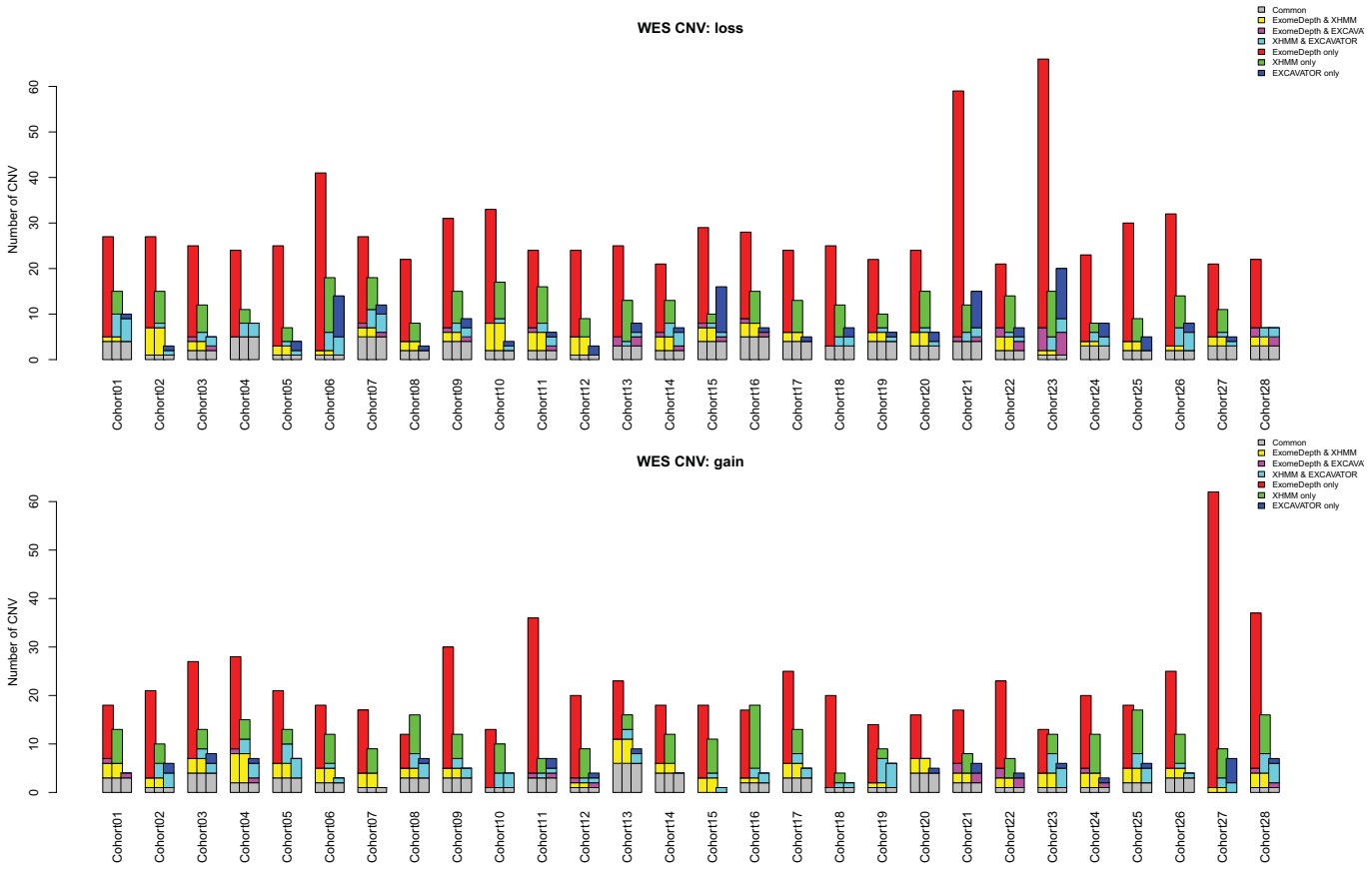

図 12　WES データから call されたサンプル毎の CNV 数．上段:loss，下段:gain．各サンプルの 3 つの縦棒は左から ExomeDepth[29]，XHMM[11]，EXCAVATOR[23] の call 数を表し，他のツールと一致した call を色分けして示した．

の rare CNV 数を図 11 に示した．ただしこれらは rare CNV と判定された call のみを対象としており，他のツールで rare CNV 以外として call された CNV は考慮していないことに注意されたい．尚，WES からの CNV call 評価方法は sensitivity のみとした．

4　比較結果

4.1　WES データからの予測結果

ExomeDepth[29]，XHMM[11]，及び EXCAVATOR[23] により call されたサンプル毎の loss 及び gain のセグメント数とセグメント長の分布を図 13 に示した．ただし ExomeDepth の最短予測長は loss 及び gain 共に 2 bp であったため，複数以上の exon を含む call と exon の半分以上を含む call のみを用いている．この結果 ExomeDepth[29] の総 call 数は loss で 1,369→802，gain で 1,035→627 となり，ExomeDepth[29] の最短予測長は loss が 2 bp，gain が 26 bp となった．一方で XHMM[11] の最短予測長は loss が 328 bp，gain が 322 bp，EXCAVATOR[23] の最短予測長は loss が 3,118 bp，gain が 4,599 bp であった．長い CNV と予測された上位を表 11 に示した．

次に今回用いた 3 つのツールの call 結果の重なりを表 12 及び図 14 に示した．ここで複数領域が互い違いに重なった場合は，3 つとも loss 若しくは gain の判定が一致して 3 つのオーバーラップ長が最も長いトリオ，2 つで共通して同じ判定でオーバーラップ長が最も長いペアの順に逐次的にカウントした．

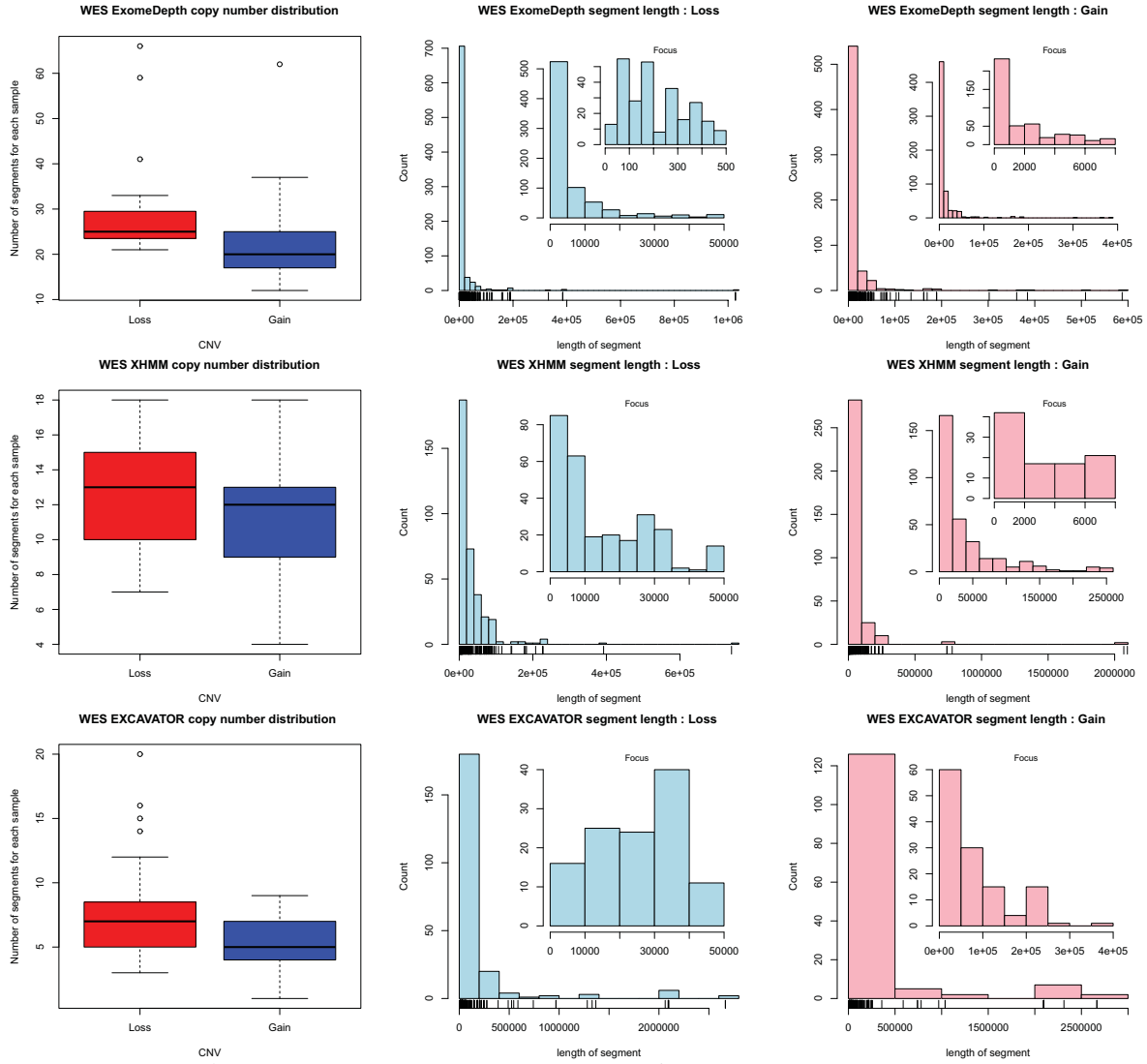

図 13　28 人の WES データから CNV を予測した結果．左図:サンプル毎の loss と gain のセグメント数の boxplot，中央図:全サンプルにおける loss の長さの分布，右図:全サンプルにおける gain の長さの分布．上段:ExomeDepth[29](比較対象 CNV のみ)，中段:XHMM[11]，下段:EXCAVATOR[23]．

4.2　Omni-CNV 予測データと WGS-CNV 予測データの比較

PennCNV[35] と ERDS[40] 及び CNVnator[1] との call の重なりを表 13 に示した．ここで loss 若しくは gain の判定が同じで 1 bp 以上重なっていれば共通と判断している．ただし表 13 の上段の 3 つの call の重なりでは，ERDS[40] と CNVnator[1] でセグメント領域が重なっている場合に，この和集合領域と PennCNV[35] のセグメントが重なっていれば 3 つ共通 call と判定した．そのため表 13 の下段の 2 つの比較結果とは一致しないことに注意されたい．また 3 つの call の重なりのベン図を図 15 に示した．Loss については CNVnator[1] の call 数が圧倒的に多く，また gain については WGS からの予測はある程度一致するが PennCNV[35] の結果とは余り一致しない．Omni チップの LRR や BAF を算出したクラスターに問題がある可能性が懸念される結果となった．

4.3　Omni-CNV 予測データと WES-CNV 予測データの比較

Omni-CNV 予測データと WES-CNV 予測データの比較結果を表 14 の上段に示した．Loss については XHMM[11] の感度が高く，EXCAVATOR[23] の陽性的中率が高い．Gain については XHMM[11] が感度，陽性的中率共に高かった．ただし Omni-CNV 予測データの gain については注意が必要である．

表 11 ExomeDepth[29] 及び XHMM[11] で 300 kbp 以上，EXCAVATOR[23] で 2 Mbp 以上と推定された CNV

Software	CNV	chr	start	end	length (bp)	Sample Name
ExomeDepth	Loss	11	100,226,849	100,558,563	331,715	Cohort26
		14	20,019,701	20,404,761	385,061	Cohort08, Cohort15, Cohort22
		14	87,387,810	88,414,222	1,026,413	Cohort23
		14	87,387,810	88,416,275	1,028,466	Cohort06
	Gain	10	37,506,624	37,890,972	384,349	Cohort27
		11	100,221,434	100,730,356	508,923	Cohort27
		12	33,592,308	34,179,850	587,543	Cohort05
		15	22,382,474	22,743,598	361,125	Cohort15
		17	43,559,804	43,861,943	302,140	Cohort24
XHMM	Loss	3	86,316,070	86,708,420	392,351	Cohort06
		10	46,961,583	47,701,761	740,179	Cohort08
	Gain	8	3,855,268	5,923,804	2,068,537	Cohort25
		10	46,961,583	47,701,761	740,179	Cohort13
		10	46,961,583	47,701,761	740,179	Cohort28
		12	33,592,253	34,371,800	779,548	Cohort05
		15	20,450,187	22,545,688	2,095,502	Cohort23
EXCAVATOR	Loss	9	68,415,102	71,080,251	2,665,150	Cohort01, Cohort07
		15	20,450,187	22,512,653	2,062,467	Cohort18
		15	20,450,187	22,541,778	2,091,592	Cohort01
		15	20,450,187	22,545,688	2,095,502	Cohort07, Cohort27
		15	20,450,187	22,546,160	2,095,974	Cohort04, Cohort11
	Gain	8	3,611,299	5,923,804	2,312,506	Cohort25
		9	68,415,102	71,080,251	2,665,150	Cohort08
		9	68,415,102	71,081,663	2,666,562	Cohort19
		15	20,450,187	22,542,982	2,092,796	Cohort19
		15	20,450,187	22,545,688	2,095,502	Cohort09, Cohort28
		15	20,450,187	22,546,160	2,095,974	Cohort08, Cohort12, Cohort23

表 12 WES 28 サンプルにおける ExomeDepth[29] (ED)，XHMM[11] (XH)，EXCAVATOR[23] (EX) の call の重なり

					Only			Total		
CNV	ED∩XH∩EX	ED∩XH	ED∩EX	XH∩EX	ED	XH	EX	ED	XH	EX
Loss	78	60	19	49	645	165	72	802	352	218
Gain	59	65	10	50	493	148	23	627	322	142

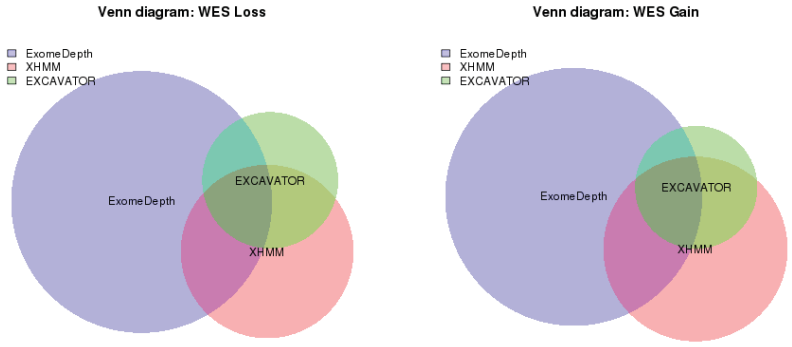

図 14 WES 28 サンプルにおける ExomeDepth[29]，XHMM[11]，EXCAVATOR[23] の call の重なり．左図:loss，右図:gain.

4.4 WGS-CNV 予測データと WES-CNV 予測データの比較

ERDS[40] 及び CNVnator[1] と WES-CNV 予測データとの比較をそれぞれ表 14 に示した．またこのベン図を図 16 に示した．尚，図 10 に示した様に ERDS[40] の比較対象領域の多くは CNVnator[1] でも call されているため，ERDS[40] と WES call のみの重なりは殆んど

表 13　Omni-CNV 予測データと WGS-CNV 予測データの CNV call の重なり．ER:ERDS[40]，CN:CNVnator[1]，PC:PennCNV[35]．

CNV	ER∩CN∩PC	ER∩CN	ER∩PC	CN∩PC	Only ER	Only CN	Only PC	Total ER	Total CN	Total PC
Loss	599	1,076	3	508	65	66,362	580	1,743	68,545	1,690
Gain	206	2,016	13	2	712	1,849	2,503	2,947	4,073	2,724

	PennCNV vs ERDS			PennCNV vs CNVnator		
CNV	Common	PennCNV	ERDS	Common	PennCNV	CNVnator
Loss	598	1,092	1,141	1,106	584	67,436
Gain	218	2,506	2,727	206	2,518	3,861

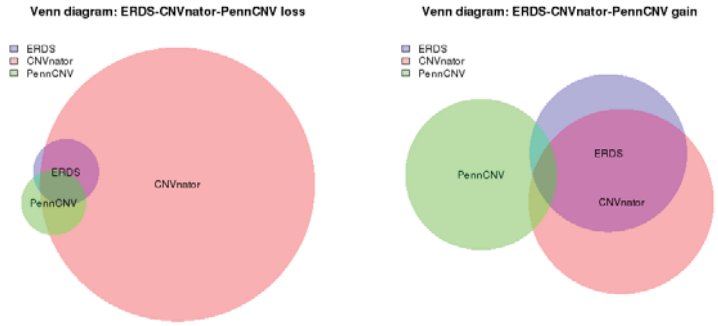

図 15　WGS-CNV 予測データと PennCNV call の重なり．左図:loss，右図:gain．ただし loss は円では表現出来ないため CNVnator のみの call 数を 66,362 から 28,000 に置き換えて表示している．

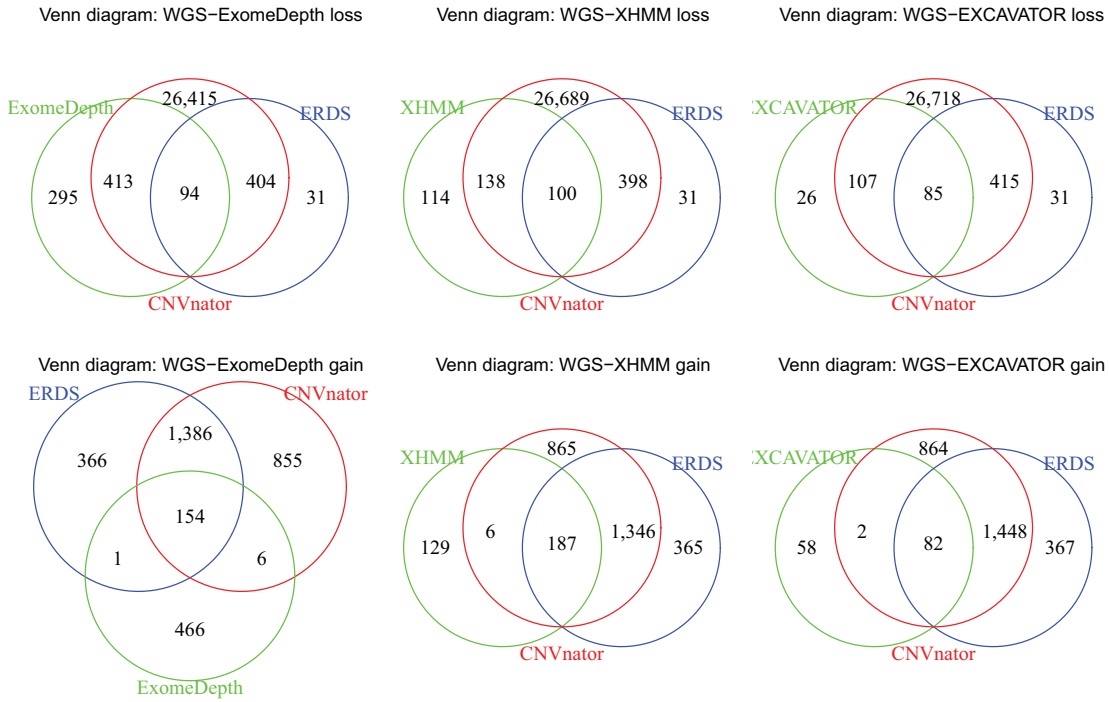

図 16　WGS-CNV 予測データと WES call の重なり (ベン図)．上段:loss，下段:gain．左図:ExomeDepth，中央図:XHMM，右図:EXCAVATOR．

0 となった．

ERDS[40] から予測された loss については，XHMM[11] の感度が高く，EXCAVATOR[23] の陽性的中率が高かった．一方で Gain については XHMM[11] が良い成績となった．

CNVnator[1] から予測された loss については，感度と陽性的中率の関係が現れているが，gain についてはやはり XHMM[11] の方が総合的に良い成績であったと言えるだろう．

表 14 PennCNV[35]，ERDS[40] 及び CNVnator[1] の推定結果を gold standard とした時の WES-CNV 予測データの推定精度. LR:loss region，NLR:not loss region，GR:gain region，NGR:not gain region.

Gold standard	WES software	Loss call				Gain call			
		call	LR	NLR	Prop.	call	GR	GLR	Prop.
PennCNV	ExomeDepth	LR	81	721	0.101	GR	59	568	0.0941
		NLR	329	-	-	NGR	2,366	-	-
		Prop.	0.198	-			0.0243	-	
	XHMM	LR	104	248	0.295	GR	71	251	0.220
		NLR	306	-	-	NGR	2,354	-	-
		Prop.	0.254	-			0.0293	-	
	EXCAVATOR	LR	79	139	0.362	GR	43	99	0.303
		NLR	331	-	-	NGR	2,382	-	-
		Prop.	0.193	-			0.0177	-	
ERDS	ExomeDepth	LR	93	709	0.116	GR	146	481	0.233
		NLR	435	-	-	NGR	1,751	-	-
		Prop.	0.176	-			0.0770	-	
	XHMM	LR	99	253	0.281	GR	187	135	0.581
		NLR	429	-	-	NGR	1,710	-	-
		Prop.	0.188	-			0.0986	-	
	EXCAVATOR	LR	83	135	0.381	GR	83	59	0.585
		NLR	445	-	-	NGR	1,814	-	-
		Prop.	0.157	-			0.0438	-	
CNVnator	ExomeDepth	LR	506	296	0.631	GR	159	468	0.254
		NLR	26,819	-	-	NGR	2,232	-	-
		Prop.	0.0185	-			0.0665	-	
	XHMM	LR	238	114	0.676	GR	190	132	0.590
		NLR	27,087	-	-	NGR	2,201	-	-
		Prop.	0.00871	-			0.0795	-	
	EXCAVATOR	LR	192	26	0.881	GR	84	58	0.592
		NLR	27,133	-	-	NGR	2,307	-	-
		Prop.	0.00703	-			0.0351	-	

表 15 Rare CNV の検出数

	ExomeDepth			XHMM			EXCAVATOR		
PennCNV loss	19	374	0.0483	29	364	0.0738	15	378	0.0382
PennCNV gain	22	196	0.101	30	188	0.138	18	200	0.0826
ERDS loss	9	1,406	0.00636	15	1,400	0.0106	4	1,411	0.00283
ERDS gain	33	395	0.0771	39	389	0.0911	17	411	0.0397
CNVnator loss	21	15,870	0.00132	24	15,867	0.00151	13	15,878	0.000818
CNVnator gain	27	7,669	0.00351	35	7,661	0.00455	16	7,680	0.00208

4.5 Rare CNV の比較

PennCNV[35]，ERDS[40]，CNVnator[1] で rare となった CNV の WES-CNV データにおける検出数を表 15 に示した. 全てで XHMM[11]，ExomeDepth[29] の順に感度が若干高い. 総 call 数は ExomeDepth[29] が最も多いにも関わらず XHMM[11] の感度が高いことは XHMM[11] の開発目的と合致した結果と言えるだろう.

5 まとめ

WES データからの CNV 予測について文献調査を行い，28 人分の実データを用いて ExomeDepth[29]，XHMM[11]，EXCAVATOR[23] の評価を行った. この結果から，

1. ExomeDepth[29] は gain の成績が余り良くなかった.
2. XHMM[11] は loss，gain 共に成績が良く，また rare CNV の検出感度も高かった.
3. EXCAVATOR[23] は感度は余り高くないが陽性

的中率は比較的高い.

との評価が今回のデータからは得られた. Renjie *et al.* (2014)[30] では ExomeDepth[29] の成績が良かったが，本データでは XHMM[11] の成績が良く，比較に用いるコントロールデータや比較の方法 (exon を跨ぐか exon の半分以上の call に限定している) に結果が大きく依存するのかも知れない. しかしながら，既報で報告されている通り，全てのツールで実用に耐えうる感度及び陽性的中率では無いと思われる.

一方で CNA 予測では染色体の短椀，長腕と言った非常に大きな領域で変化していることも多く，WES データからでもある程度の精度で予測出来るかもしれない. しかしながら実際の固形がんの臨床検体では白血球や血管と言ったがんで無い細胞の混入が通常避けられないことから，CNA 予測ではがん細胞と非がん細胞が混ざっていることを前提に予測方法を構築する必要がある. 今回調査した WES データからの CNV 予測方法の中にはがんの CNA 予測を適用範囲とするソフトウエアもあったが，複数種類の細胞集合への適用を述べている予測法は ExomeCNV[32] のみであった. 尚，Illumina の SNP チップからの CNA 予測では ASCAT[28] や GPHMM[17] らがこれをモデル化している.

謝辞：本稿は，文部科学省 社会システム改革と研究開発の一体的推進における「大規模分子疫学コホート研究の推進と統合」の「B. ゲノム解析」で行わ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.

引 用 文 献

1) Alexej Abyzov, Alexander E. Urban, Michael Snyder, and Mark Gerstein. Cnvnator: An approach to discover, genotype, and characterize typical and atypical cnvs from family and population genome sequencing. *Genome Research*, Vol. 21, No. 6, pp. 974–984, 2011.

2) Can Alkan, Jeffrey M Kidd, Tomas Marques-Bonet, Gozde Aksay, Francesca Antonacci, Fereydoun Hormozdiari, Jacob O Kitzman, Carl Baker, Maika Malig, Onur Mutlu, S Cenk Sahinalp, Richard A Gibbs, and Evan E Eichler. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genetics*, Vol. 41, No. (10), pp. 1061–1067, Oct 2009.

3) Matteo Benelli, Giuseppina Marseglia, Genni Nannetti, Roberta Paravidino, Federico Zara, Franca Dagna Bricarelli, Francesca Torricelli, and Alberto Magi. A very fast and accurate method for calling aberrations in array-cgh data. *Biostatistics*, Vol. 11, No. 3, pp. 515–518, 2010.

4) Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, Vol. 57, No. (1), pp. 289–300, 1995.

5) Valentina Boeva, Andrei Zinovyev, Kevin Bleakley, Jean-Philippe Vert, Isabelle Janoueix-Lerosey, Olivier Delattre, and Emmanuel Barillot. Control-free calling of copy number alterations in deep-sequencing data using gc-content normalization. *Bioinformatics*, Vol. 27, No. 2, pp. 268–269, 2011.

6) Lachlan J M Coin, Julian E Asher, Robin G Walters, Julia S El-Sayed Moustafa, Adam J de Smith, Rob Sladek, David J Balding, Philippe Froguel, and Alexandra I F Blakemore. cnvhap: an integrative population and haplotype-based multiplatform model of snps and cnvs. *Nat Methods*, Vol. 7, No. (7), pp. 541–546, Jul 2010.

7) Lachlan J.M. Coin, Dandan Cao, Jingjing Ren, Xianbo Zuo, Liangdan Sun, Sen Yang, Xuejun Zhang, Yong Cui, Yingrui Li, Xin Jin, and Jun Wang. An exome sequencing pipeline for identifying and genotyping common cnvs associated with disease with application to psoriasis. *Bioinformatics*, Vol. 28, No. 18, pp. i370–i374, 2012.

8) International HapMap Consortium. Integrating ethics and science in the international hapmap project. *Nature Revivew Genetics*, Vol. 5, No. (6), pp. 467–475, Jun 2004.

9) The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, Vol. 467, pp. 1061–1073, Oct 2010.

10) Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AWC, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Wellcome Trust Case Control Consortium, Tyler-Smith C, Carter NP, Lee C, Scherer SW, and Hurles ME. Origins and functional impact of copy number variation in the human genome. *Nature*, Vol. 464, pp. 704–712, April 2010.

11) Menachem Fromer, Jennifer L. Moran, Kimberly Chambert, Eric Banks, Sarah E. Bergen, Douglas M. Ruderfer, Robert E. Handsaker, Steven A. McCarroll, Michael C. O'Donovan, Michael J. Owen, George Kirov, Patrick F. Sullivan, Christina M. Hultman, Pamela Sklar, and Shaun M. Purcell. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *The American Journal of Human Genetics*, Vol. 91, No. 4, pp. 597 – 607, 2012.

12) Faraz Hach, Fereydoun Hormozdiari, Can Alkan, Farhad Hormozdiari, Inanc Birol, Evan E Eichler, and S Cenk Sahinalp. mrsfast: a cache-oblivious algorithm for short-read mapping. *Nature Methods*, Vol. 7, No. (8), pp. 576–577, Aug 2010.

13) Günter Klambauer, Karin Schwarzbauer, Andreas Mayr, Djork-Arnè Clevert, Andreas Mitterecker, Ulrich Bodenhofer, and Sepp Hochreiter. cn.mops: mixture of poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Research*, Vol. 40, No. 9, p. e69, 2012.

14) Ryan Koehler, Hadar Issac, Nicole Cloonan, and Sean M. Grimmond. The uniqueome: a mappability resource for short-tag sequencing. *Bioinformatics*, Vol. 27, No. 2, pp. 272–274, 2011.

15) Niklas Krumm, Peter H. Sudmant, Arthur Ko, Brian J. O'Roak, Maika Malig, Bradley P. Coe, NHLBI Exome Sequencing Project, Aaron R. Quinlan, Deborah A. Nickerson, and Evan E. Eichler. Copy number variation detection and genotyping from exome sequence data. *Genome Research*, Vol. 22, No. 8, pp. 1525–1532, 2012.

16) Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol.*, Vol. 10, No. (3), p. R25, Mar 2009.

17) 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.

18) Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, Vol. 25, No. (14), pp. 1754–60, Jul 2009.

19) Jason Li, Richard Lupat, Kaushalya C. Amarasinghe, Ella R. Thompson, Maria A. Doyle, Georgina L. Ryland, Richard W. Tothill, Saman K. Halgamuge, Ian G. Campbell, and Kylie L. Gorringe. Contra: copy number analysis for targeted resequencing. *Bioinformatics*, Vol. 28, No. 10, pp. 1307–1313, 2012.

20) Ruiqiang Li, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming Yiu, Karsten Kristiansen, and Jun Wan. Soap2: an improved ultrafast tool for short read alignment. *Bioinformatics*, Vol. 25, No. (15), pp. 1966–7, Aug 2009.

21) DePristo M., Banks E., Poplin R., Garimella K., Maguire J., Hartl C., Philippakis A., del Angel G., Rivas MA., Hanna M., McKenna A., Fennell T., Kernytsky A., Sivachenko A., Cibulskis K., Gabriel S., Altshuler D, and Daly M. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature Genetics*, Vol. 43, pp. 491–498, 2011.

22) Alberto Magi, Matteo Benelli, Seungtai Yoon, Franco Roviello, and Francesca Torricelli. Detecting common copy number variants in high-throughput sequencing data by using jointslm algorithm. *Nucleic Acids Research*, Vol. 39, No. 10, p. e65, 2011.

23) Alberto Magi, Lorenzo Tattini, Ingrid Cifola, Romina D'Aurizio, Matteo Benelli, Eleonora Mangano, Cristina Battaglia, Elena Bonora, Ants Kurg, Marco Seri, Pamela Magini, Betti Giusti, Giovanni Romeo, Tommaso Pippucci, Gianluca De Bellis, Rosanna Abbate, and Gian Franco Gensini. Excavator: detecting copy number variants from whole-exome sequencing data. *Genome Biology*, Vol. 14, No. 10, p. R120, 2013.

24) J. C. Marioni, N. P. Thorne, and S. Tavaré. Biohmm: a heterogeneous hidden markov model for segmenting array cgh data. *Bioinformatics*, Vol. 22, No. 9, pp. 1144–1146, 2006.

25) Love MI, Myšičkovà A, Sun R, Kalscheuer V, Vingron M, and Haas SA. Modeling read counts for cnv detection in exome sequencing data. *Stat Appl Genet Mol Biol.*, Vol. 8, No. 10(1), Nov 2011.

26) Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods*, Vol. 5, No. (7), pp. 621–8, Jul 2008.

27) A. B. Olshen and E. S. Venkatraman. Circular binary segmentation for the analysis of arraybased dna copy number data. *Bioinformatics*, Vol. 5, pp. 557–572, 2004.

28) 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.

29) Vincent Plagnol, James Curtis, Michael Epstein, Kin Y. Mok, Emma Stebbings, Sofia Grigoriadou, Nicholas W. Wood, Sophie Hambleton, Siobhan O. Burns, Adrian J. Thrasher, Dinakantha Kumararatne, Rainer Doffinger, and Sergey Nejentsev. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*, Vol. 28, No. 21, pp. 2747–2754, 2012.

30) Renjie Tan, Yodong Wang, Sarah E. Kleinstein, Yongzhuang Liu, Xiaolin Zhu, Hongzhe Guo, Qinghua Jiang, Andrew S. Allen, Mingfu Zhu. An evaluation of copy number variation detection tools from whole-exome sequencing data. *Human Mutation*, Mar 2014.

31) McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PIW, Maller JB, Kirby A, Elliott AL, Parkin M, Hubbell E, Webster T, Mei R, Veitch J, Collins PJ, Handsaker R, Lincoln S, Nizzari M, Blume J, Jones KW, Rava R, Daly MJ, Gabriel SB, and Altshuler D. Integrated detection and population-genetic analysis of snps and copy number variation. *Nat Genet*, Vol. 40, pp. 1166–1174, 2008.

32) Jarupon Fah Sathirapongsasuti, Hane Lee, Basil A. J. Horst, Georg Brunner, Alistair J. Cochran, Scott Binder, John Quackenbush, and Stanley F. Nelson. Exome sequencing-based copy-number variation and loss of heterozygosity detection: Exomecnv. *Bioinformatics*, Vol. 27, No. 19, pp. 2648–2654, 2011.

33) 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.

34) E. S. Venkatraman and Adam B. Olshen. A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics*, Vol. 23, No. 6, pp. 657–663, 2007.

35) 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.

36) David Weese, Anne-Katrin Emde, Tobias Rausch, Andreas Döring, and Knut Reinert. Razers-fast read mapping with sensitivity control. *Genome Research*, Vol. 19, No. 9, pp. 1646–1654, 2009.

37) Chao Xie and Martti Tammi. Cnv-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, Vol. 10, No. 1, p. 80, 2009.

38) Yan Guo, Quanghu Sheng, David C. Samuels, Brian Lehmann, Joshua A. Bauer, Jennifer Pietenpol and Yu Shyr. Comparative study of exome copy number variation estimation tools using array comparative genomic hybridization as control. *BioMed Research International*, Vol. 2013, , Nov 2013.

39) Seungtai Yoon, Zhenyu Xuan, Vladimir Makarov, Kenny Ye, and Jonathan Sebat. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research*, Vol. 19, No. 9, pp. 1586–1592, 2009.

40) Mingfu Zhu, Anna C. Need, Yujun Han, Dongliang Ge, Jessica M. Maia, Qianqian Zhu, Erin L. Heinzen, Elizabeth T. Cirulli, Kimberly Pelak, Min He, Elizabeth K. Ruzzo, Curtis Gumbs, Abanish Singh, Sheng Feng, Kevin V. Shianna, and David B. Goldstein. Using erds to infer copy-number variants in high-coverage genomes. *The American Journal of Human Genetics*, Vol. 91, No. 3, pp. 408–421, 2012.