2007－BIO－9 (1)
2007／6／14

高密度SNPチップを用いたコピー数解析における データ精製と閾値の関係について

小長谷 明彦，長谷川 亜樹，小島俊男

独立行政法人 理化学研究所 ゲノム科学総合研究センター

アブストラクト

ヒトゲノムにおける高密度 SNP チップを用いたコピー数データの解析は、様々な遺伝性疾患の原因となる遺伝子の特定や診断を行う上で有用な情報を取り出す有効な手段となりつつある。しかしながら、実験に用いた全ての高密度 SNP チップが高いシグナル／ノイズ比を示すわけではなく、実際のコピー数変異の発見には、適切なデータ精製と閾値の設定が不可欠である。本稿ではコピー数判定のために設けたシグナル強度の閾値が、コピー数変異発見に及ぼす影響について考察する。

1 はじめに

人のゲノムには、５００塩基から１０００塩基毎にSNPと呼ばれる単一塩基多型（Single Nucleotide Polymorphism)があり、個人差をもたらす要因の一つといわれている。SNPはアミノ酸の変異や遺伝子発現の変異を引き起こし、結果的に各個人の代謝の違いやシグナル伝達の違いを引き起こす。ハンチントン病や血友病のような遺伝病や血液型あるいはアルコール代謝能力の違いなどもSNPの違いに由来する。国際HAPMAP計画を中心に、ヒトの持つ変異の網羅的な収集が進められ、これまでに、４００万個以上のSNPがインターネット上で公開されている[1]。

ゲノム変異の研究の進行により、ゲノム上にはSNPだけでなく、コピー数多型（Copy Number Polymorphism: CNP）と呼ばれる変異があることがあきらかになってきた[Redon2006]。CNPはDNA領域そのものが数千塩基対から数万塩基対という長い領域にわたって減少(loss)、過剰(gain)、増幅(amplification)することにより、遺伝子のコピー数が変化する現象である。ガン細胞に関しては、これまでにも、染色体レベルおよびバンドレベルでの大きな変異が存在することが、ゲノムワイドなハイブリダイゼーションによる解析（Comparative Genome Hybridization(CGH 法)）で知られていた[Jain2001]。遺伝子コピー数を変える様な変異（Copy Number Variation:CNV）は生殖系細胞においても生じており、疾病との関連性が強く示唆されている[kojima2006]。近年、遺伝子数の変化は健常人にも存在することが判明し、体系的なコピー数多型(CNP)の解析方法が求められている[Komura2006]。

ゲノム上で生じた変異に関するもっとも確実な判定法は再シークエンスである。アセンブリされた複数のゲノム配列があれば、配列を比較することで、不一致（mismatch)、非整合（unmatch)、コピー数異常(copy unmatch)、逆位(inversion)などの変異を検出することができる[Khaja2006]。配列シークエンス技術の向上は著しく、あと数年もすれば個人ゲノム配列から個人単位の変異（ＳＮＰやＣＮＰ）を同定することも技術的には可能な時代にはいる。

比較的安価で、現時点でもっとも大量の情報が得られるのは、高密度SNPチップを用いたコピー数変異解析法である[Kojima06, Komura06, Redon06]。最新の高密度SNPには、数十万個以上のプローブが搭載されており、全ゲノム領域を5K塩基から１０K塩基間隔でカバーしている。これにより、標識し

[1] http://www.hapmap.org/

た DNA 断片を染色体標本にハイブリダイズさせる Comparative Genome Hybridization(CGH 法)[Kallioniemi1992]や、マイクロアレイ上に配置した DNA 断片にハイブリダイズさせるアレイ CGH 法 [Albertson2003]等と比べ、はるかに高い解像度でコピー数多型解析を行うことが期待できる。

しかしながら、高密度 SNP チップのシグナル強度は PCR およびハイブリダイゼーションの影響を受けるため、多くのノイズを含んでいる。特に、高密度化により、スポット毎のシグナル強度は非常に微弱なものとなっており、SNP に完全に照合するプローブ（Perfect Match:PM）のシグナル強度と不完全に照合するプローブ(Miss Match:MM)との差は小さい。このため、高密度 SNP チップを用いたコピー数多型解析においては、減少(loss)・増加(gain)の有無を判定するための閾値の設定が極めて重要な意味を持つ。本稿では、閾値の設定がデータマイニング手法によるコピー数変異部位の発見におよぼす影響について考察する。具体的にはインターネット上で公開済みのデータセット[2]において、減少、増加を判定するための閾値を変動させたときに、既知の減少・増加部位（ＣＮＰ領域）を正例としたときの感受性（sensitivity）について考察する。

本稿の構成は以下の通りである。はじめに、2 節において遺伝子コピー数多型について実例をあげて説明する。3 節では、本稿で述べる手法で使用された高密度ＳＮＰチップの仕様について説明する。次に、4 節において作成したＣＮＰ解析システム VARSearch の概要を、5 節において、ＣＮＰの検索結果と閾値の影響について報告する。最後に、6 節においてまとめと今後の展望について述べる。

2　遺伝子コピー数多型（Copy Number Polymorphism: CNP）解析

DNA には単一塩基多型（SNP）のほかに、重複や欠失から生じたコピー数による多型（CNP）があり、疾患および多様性の要因の一つとなっている。CNP は数千塩基対から数百万塩基対という大きな領域そのものが重複、欠失したものであり、遺伝子領域を含む場合は遺伝子の数そのものの数が異なるという問題を生じる。実験により確認されたコピー数変異の例を図１に示す。この患者では染色体の一方に４．５Ｍ塩基対にわたる欠損が報告されている。この領域には１８個の遺伝子が含まれており、そのコピー数が通常の半分になっていることとてんかん発作などの疾病との関連が示唆されている[Kojima2006]。

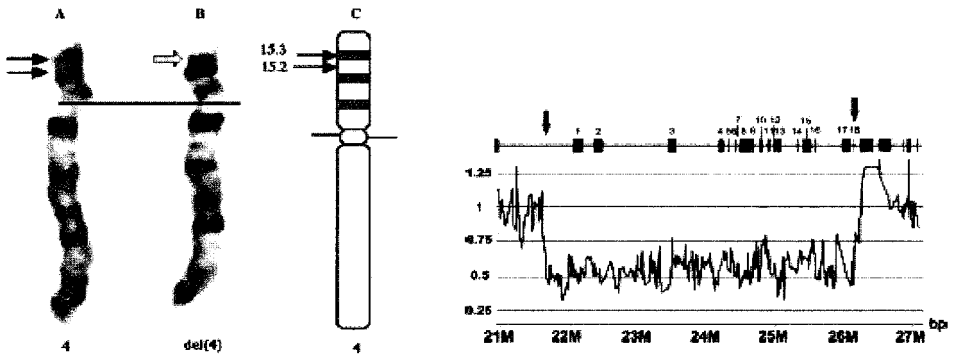

図１　実験により確認されたコピー数変異の例（文献[Kojima2006]より複製）

[2] http://projects.tcag.ca/variation/

このようなコピー数変異(CNV)あるいはコピー数多型(CNP)を同定するために、Comparative Genome Hybridization(CGH 法)、アレイ CGH 法、SNP チップ法などがこれまでに提案され、疾患解析に利用されている。

CGH 法は１９９２年に Kallioniemi らにより提案されたゲノムワイドなハイブリダイゼーションによる DNA 配列のコピー数の解析法である[Kallioniemi1992]。これまでに、癌細胞など、染色体レベルおよびバンドレベルでの大きな減少(loss)、過剰(gain)、増幅(amplification)の解析などに利用されてきた[Jain2001]。CGH 法では、対象となる腫瘍組織の DNA と正常組織の DNA を異なる蛍光色素で標識し、有糸分裂期の染色体標本と競合的なハイブリダイゼーションを行う。現象や増幅のある部位は、ハイブリダイズする DNA の量に偏りがでるため、蛍光顕微鏡などにより識別できる。CGH 法は技術的に確立しており、これまでに固形腫瘍を中心にゲノム異常解析法として広く利用されている。しかしながら、１コピーの増幅の検出には１M 塩基対以上、１コピーの減少の検出には、少なくとも１０M 塩基対以上の大きさの変異でないと検出できないなど、分解能に問題がある。

アレイ CGH 法では、CGH 法で用いた染色体標本の代わりに、多数のクローン化 DNA 断片を配置したマイクロアレイを使うことにより、解像度を大幅に向上している。BAC クローンを用いたマイクロアレイでは、およそ１０万塩基対の解像度でゲノムワイドに DNA コピー数多型を検出することが可能である[Albertson2003]。ただし、疾患関連遺伝子を具体的に同定するためには１０万塩基対の解像度ではまだまだ不十分であり、より解像度の高い方法が求められている。

SNP チップ法は、SNP タイピング用に作成された高密度 SNP チップを利用したコピー数多型解析である[kojima2006, Komura2006, Redon2006]。最新の高密度 SNP チップは、およそ５０万箇所の SNP のタイプを検出するように設計されており、ゲノムのほぼ全域を５千-１万塩基対毎に一つの割合で配置されている（図２）。SNP 部位にコピー数変異がある場合、シグナル値がコピー数分だけ増加あるいは減少するため、その変動を検出することにより変異を予測することが可能となる。

3　高密度 SNP チップ

本稿で紹介する解析に用いた SNP データは、HAPMAP 計画でインターネット上に公開した 270 名分の HAPMAP コレクションである[3]。このデータは GeneChip 500K Mapping Array Set[4]の early access(EA)版を用いている。本チップセットは、使用する制限酵素の違いにより、２つのアレイ（250K Sty, 250K Nsp）から構成される。各アレイには、約２５万個の SNP を識別するためのプローブセットが用意され、各プローブセットは２５塩基の長さを持つ２４本のプローブから構成される。これらは、33 塩基の仮想プローブ領域を認識するように設計されており、それぞれの領域のアレルタイプと完全に照合する６対の Perfect Match（PA および PB）と不完全に照合する６対の Miss Match(MA および MB)のプローブから構成される（図３）。したがって、チップセット全体のプローブ数は約１２００万本となる。

GeneChip でのシグナル強度の測定は以下の手順で行われる[5]。測定したい組織または細胞から DNA を抽出し、制限酵素 StyI または NspI を用いて DNA を分断する。PCR により DNA 断片を 200 から 1100 塩基の範囲で増幅させ、分解試薬を用いて細かく分断する。断片 DNA 末端をラベル標識し、GeneChip とハイブリダイズする。Chip を洗浄し、染色し、蛍光強度をスキャニングする。スキャニングした画像データからスポット位置を割り出し、蛍光強度を平均化してスポットのシグナル強度とする。上記の処理により得られたデータが.cel に格納されている。

---

[3] http://130.14.29.110/projects/geo/query/acc.cgi?acc=GSE5013
[4] http://www.affymetrix.com/support/technical/byproduct.affx?product=500k
[5] https://www.affymetrix.com/support/downloads/manuals/500k_assay_manual.pdf

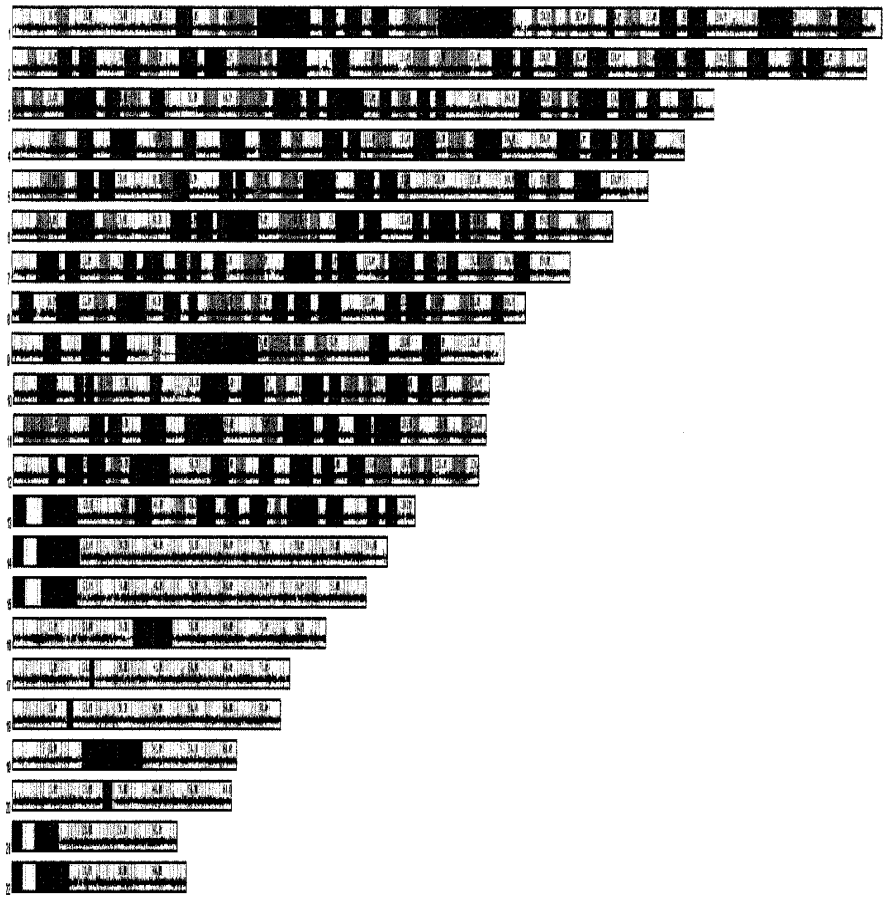

図２　高密度ＳＮＰチップ上でのプローブの分布

GeneChip Probe Information - Microsoft Internet Explorer

rs4308923

LOCATION	chr: 1, band: , pos: 24150645								
FRAGMENT	len: 468, start: 24150262, stop: 24150729								
ALLELE	A: A, B: T								
FREQUENCY	A: 0.58, B: 0.42, H: 0.48								
CHIP	EA500k_NspI								
GC CONTENTS	A:123G:96T:147C:96N:0								
FLANK	tatcaggaaacaaaga.gtcacttgaagcatta								
Q1	==========*==========..... (-4) sense	PA probe[t]	PA target[a]	PB probe[t]	PB target[a]	MA probe[a]	MA target[a]	MB probe[a]	MB target[a]
Q2	..==========*==========..... (-2) antisense	PA probe[g]	PA target[c]	PB probe[g]	PB target[c]	MA probe[c]	MA target[c]	MB probe[c]	MB target[c]
Q3	....==========*==========.... (+0) sense	PA probe[t]	PA target[a]	PB probe[a]	PB target[t]	MA probe[c]	MA target[a]	MB probe[g]	MB target[t]
Q4	....==========*==========.... (+0) antisense	PA probe[a]	PA target[t]	PB probe[t]	PB target[a]	MA probe[g]	MA target[t]	MB probe[c]	MB target[a]
Q5	.....==========*==========... (+1) sense	PA probe[c]	PA target[g]	PB probe[c]	PB target[g]	MA probe[g]	MA target[g]	MB probe[g]	MB target[g]
Q6	.....==========*========== (+4) sense	PA probe[t]	PA target[a]	PB probe[t]	PB target[a]	MA probe[a]	MA target[a]	MB probe[a]	MB target[a]

図３　高密度ＳＮＰチップで使用されているプローブ情報の例

シグナル値が得られてからの後処理としては、外れ値の除去、全体分布の正規化、ノイズの除去などを行う。GeneChip のプローブは基本的には DNA 中でユニークとなるように設計されているが、実際には重複部位が多数存在する。GeneChip 500KEA には２００箇所以上ヒットするプローブもあり、高いシグナル強度を示す。このようなスポットは解析時にアーティファクトとなるのであらかじめ除外する必要がある。一方、重複部位があっても DNA 断片の長さの関係から PCR で複製されないものもあるので注意が必要である。

正規化に関しては、LOWESS 法[Yang2002]、Quantile 法[Bolstad2003]、対数正規分布法[Konishi2005]など様々な手法が提案されており、それぞれ利点、欠点があり、目的に応じて使い分ける必要がある。SNP チップの場合、遺伝子発現チップと異なりダイナミックレンジが小さく、比較的低いシグナル値に集中するという特性を持つ。また、PM と MM との差はわずかであり、特に、SNP と対応していない場合の PM は MM と似たようなシグナル特性を持つ。

生化学実験においては様々な段階でデータエラーが混在する可能性がある。GeneChip の場合{PA、PB、MA、MB}を一組とし、これらを６セット用意することにより冗長性を持たしている。各セットはチップ上で分散して配置されているため、チップの一部が汚染されていても、汚染されたセットを除去することでプローブのシグナル値を正しく求めることができる。

SNP においては、アレル型を A、B とすると、遺伝子型としては AA、AB、BB の３種がある。PA が A、PB が B を認識するプローブとして設計されているとすると、遺伝子型が AA のときは PA＞PB かつ PB≒MA≒MB、AB のときは PA=PB かつ PA>MA かつ PA>MB、BB のときは PA<PB かつ PA≒MA≒MB となる。もし、この SNP において DNA の減少が起きていると、遺伝子型は AX または BX となり、それぞれ、PA または PB が通常の半分かつ PB≒MB または PA≒MA となる。逆に、増幅により AAB または ABB になった場合は、PA または PB が通常よりも強くなる。

４　コピー数多型解析環境 VARSearch

高密度 SNP チップから得られたシグナル強度からコピー数多型を推定するための解析環境 VARsearch を構築した。VARSearch は実験データおよび解析データを保持するためのデータベースと、解析エンジンおよび表示系の３層モデルからなる。データベースには MySQL を用い、高速応答を実現するためにクラスタ化されている。解析エンジンは大量データ処理のためのバッチシステムと検索用のオンラインシステムからなる。オンラインシステムは検索条件に合わせて動的なデータ検索ならびに加工を行い、検索結果を視覚化してブラウザー上に表示する。表示系には、標準的なブラウザーを用いる。

５　結果と考察

VARSearch により解析した結果を図４に示す。図中で、数字は染色体番号、横の筋は各チップのシグナル強度の標準データからの割合を示す。各プローブにおいて標準よりも高い場合は赤で低い場合は青で表示している。すなわち、赤く表示されている部位は増幅 CNP 部位、青く表示されている部位は減少 CNP 部位の候補となっている。どの部位が CNP 候補となるかは、対象とする組織、実験データの精度、使用する標準データセット、データ精製方法、増幅または減少と判定するための閾値の与え方に強く依存する。癌細胞のような大きな変異が普遍的にみられる場合には閾値は高めに設定してもはっきりと変異部位を確認できるが、体細胞において CNP 部位を同定するためには、生化学実験およびデータ加工法から生じるノイズとの区別が重要となる。

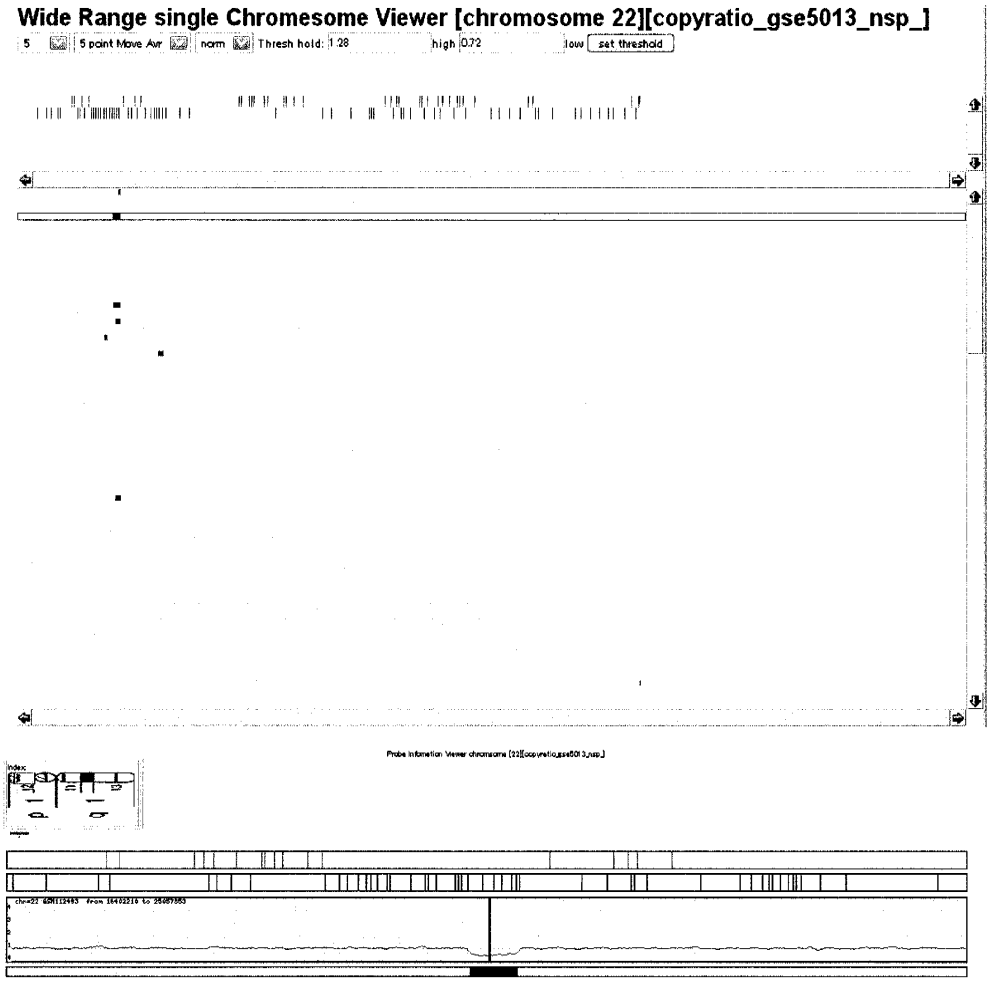

図４　検出された CNP の例：赤は増加、青は欠失をあらわす

図５および図６に、２２番染色体において既知の CNP を基準として、2 倍体とみなす（CNP でないとする）コピー数の閾値を変化させたときの感受性(sensitivity)の変化および真陽性の数の変化を示す。CNP 既知領域はＳＮＰの個数に比べ圧倒的に少ないので特異性（specificity）はほとんど変化しない（データ無し)。CNP 領域において対応するＤＮＡが１本以下の場合（欠失)、3 本以上の場合（増加)、それぞれ期待されるコピー数は０．５以下および１．５以上となるはずであるが、そのような明確な差異がでることはまれである。本当に変異のある CNP 領域を見落とさないためには、感受性を大きく下げない範囲でできるだけ多くの CNP 既知領域を覆うような閾値が必要となる。

図５が示すように、2 倍体とみなす（CNP でないとする）コピー数の閾値の上限（ub）は 1.3 まで下げても感受性はほとんど変化しない。一方、下限(lb)は 0.5 から徐々に下がりだす。ただし、絶対数でみると 0.7 付近から増えだす。このことから、閾値は 0.7 から 1.3 近辺に設定すれば取りこぼしは少ないと判断される。

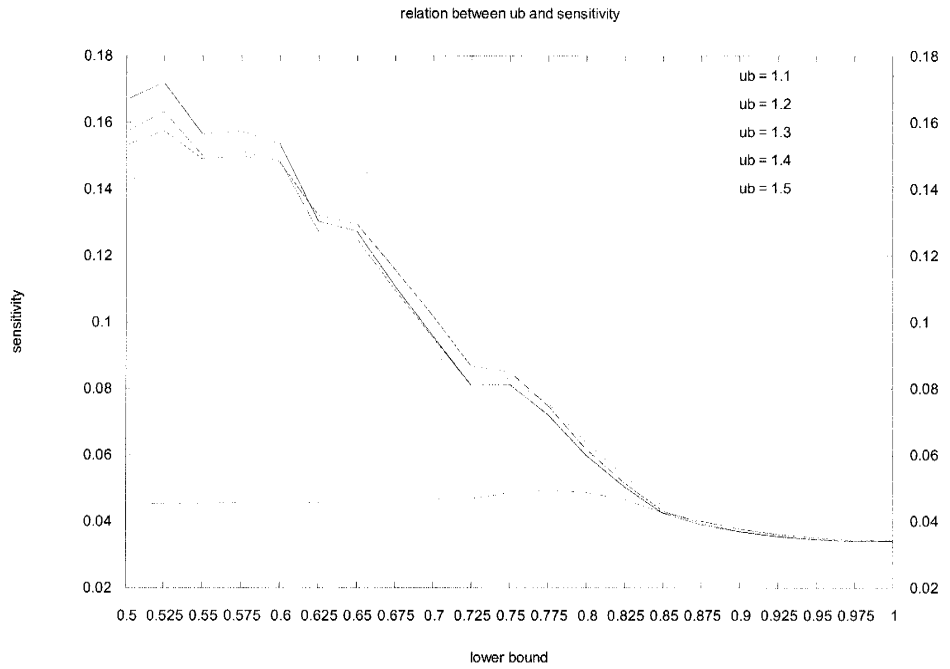

図５　閾値の設定に関する感受性と特異性との関係

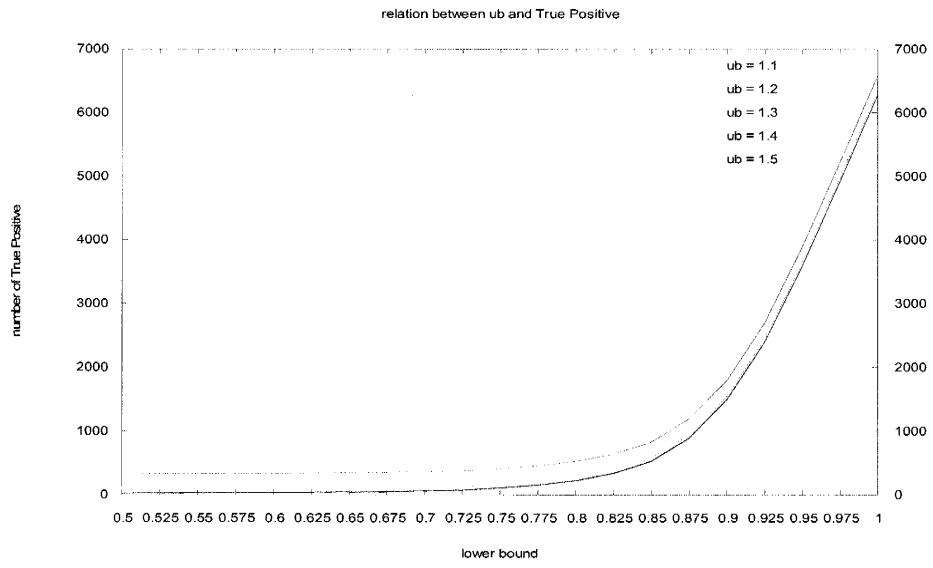

図６　閾値を変化させたときに検出された既知ＣＮＰ部位の数の変化

6　まとめ

国際 HAPMAP 計画コレクションから発表された GeneChip データを元に、VARSearch システムを用いた際の CNP の予測結果について述べた。既知の CNP 領域を標本とした感受性テストでは、2 倍体（標準）とみなすコピー数の範囲を 0.7 から 1.3 とすることで感受性の感度を落さずに対象となる変異の数を増やせることを確認した。今後の課題としては、感受性をさらに高めるためのデータ精製手法の開発があげられる。

謝辞　VARSearch の表示システムを作成して頂いた井出卓宏氏（エスカシステム）、大木真吾氏（理化学研究所）、特異性の解析をして頂いた田中彰氏（東工大）に感謝いたします。

参考文献

[Kallioniemi1992] Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D, Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors, Science, Oct 30;258(5083) (1992) 818-21.

[Alverton2003]　D.G.Albertson and D.Pinkel, Genomic microarrays in human genetic disease and cancer, Hum. Mol. Genet. 12 (2003) R145-R152

[Jain2001] A.N.Jain, K.Chin, A.L.Borresen-Dale, B.K.Erikstein, P.E.Lonning, R.Kaaresen, and J.W.Gray, Quantitative analysis of chromosomal CGH in human breast tumors associates copy number abnormalities with p53 status and c patient survival, Proc. Natl. Acad. Sci. USA 98 (2001) 7952-7957

[Veltman2002]　J.A.Veltman, E.F.Schoenmakers, B.H.Eussen, I.Janssen, G.Merkx, B.van Cleef, C.M.van Ravenswaaij, H.G.Brunner, D.Smeets, and A.G.van Kessel, High-throughput analysis of subtelomeric chromosome rearrangements by use of array-based comparative genomic hybridization, Am. J. Hum.Genet. 70 (2002) 1269-1276

[Kojima2006]　T.Kojima, W.Mukai, D.Fuma, Y.Ueda, M.Okada, Y.Sakaki, and S.Kaneko, Determination of genomic breakpoints in an epileptic patient using genotyping array, Biochem. Biophys. Res. Commun. 341 (2006) 792-796

[Iafrate2004]　A.J.Iafrate, L.Feuk, M.N.Rivera, M.L.Listewnik, P.K.Donahoe, Y.Qi, S.W.Scherer, and C.Lee, Detection of large-scale variation in the human genome, Nat.Genet. 36 (2004) 949-951

[Sebat2004]　J.Sebat, B.Lakshmi, J.Troge, J.Alexander, J.Young, P.Lundin, S.Maner, H.Massa, M.Walker, M.Chi, N.Navin, R.Lucito, J.Healy, J.Hicks, K.Ye, A.Reiner, T.C.Gilliam, B.Trask, N.Patterson, A.Zetterberg, and M.Wigler, Large-Scale Copy Number Polymorphism in the Human Genome, Science 305 (2004) 525-528

[Khaja2006]　R.Khaja, J.Zhang, J.R.MacDonald, Y.He, A.M.Joseph-George, J.Wei, M.A.Rafiq, C.Qian, M.Shago, L.Pantano, H.Aburatani, K.Jones, R.Redon, M.Hurles, L.Armengol, X.Estivill, R.J.Mural, C.Lee, S.W.Scherer, and L.Feuk, Genome assembly comparison identifies structural variants in the human genome, Nat. Genet. 38 (2006) 1413-1418

[Komura2006]　D.Komura, F.Shen, S.Ishikawa, K.R.Fitch, W.Chen, J.Zhang, G.Liu, S.Ihara, H.Nakamura, M.E.Hurles, C.Lee, S.W.Scherer, K.W.Jones, M.H.Shapero, J.Huang, and H.Aburatani, Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays, Genome Res. 16 (2006) 1575-1584

p [Redon2006] R.Redon et al, Global variation in copy number in the human genome, Nature 444 (2006) 444-454

[Yang2002] Yang YH, Dudoit S, Luu P, M.Lin D,Peng V, Ngai J, Speed TP, Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, Nucleic Acids Res., 30, e15 (2002)

[Bolstad2003] Bolstad BM, Irizarry RA, Astrand M and Speed TP, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias,　Bioinformatics, 19, (2003) 185-193

[Konishi2005] Konishi T, A thermodynamic model of transcriptome formation, Nucleic Acids Res., 33, (2005) 6587–6592