移動平均乖離度を利用したコピー数多型領域発見手法の開発

鈴木千絵　(指導教員：瀬々潤)

1　はじめに

コピー数多型 (Copy Number Variation; CNV) と疾患の関連は長く調べられてきた．代表例としてダウン症があり，21 番染色体全域に渡るコピー数多型が起こっている．近年，より詳細な CNV と疾患の関連が研究されており，自閉症，統合失調症などの遺伝要因が強いと思われている疾患．更には，がん細胞においてもゲノムの一部で反復回数が異なる部位が発見され，CNV の関与が示唆されている．CNV と疾患の対応が急激に分かるようになってきた要因として，超並列シーケンサ（次世代シーケンサ; NGS）の発展がある．CNV を含む患者のゲノム配列を NGS で読んだ場合，CNV 領域が他の領域に比べ多く読まれるあるいは，少なく読まれると考えられる．一方で，NGS によるゲノム解読は必ずしも全域に渡って均一に読まれるわけではなく，領域毎に激しく増減することも多く，CNV 領域を見つけることは容易ではない．本研究では，全ゲノム領域に対する高精度な CNV 検出手法の開発に向け，着目した位置における狭域の平均リード数と，より広域における平均リード数を比較し，その乖離度を求め，差が大きな位置でコピー数の変化が起こったことを検出する手法を開発した．

2　関連研究

NGS は，断片化された DNA 配列を大量に読み取る機器である．DNA 配列の解読価格は，最初のヒトゲノム解読が行われた 15 年前に比べ約 10 万分の 1 のコストで可能となっており，世界中でヒトのゲノム配列解読が進んでいる [1]．その中でも，疾患関連因子の探索は急速な進展を見せており GWAS catalog[2] や The Cancer Genome Atlas[3] などに，ゲノム配列と疾患の対応が蓄積されている．

更なる疾患とゲノム配列の対応を理解するために，NGS で読まれた全ゲノム配列から，CNV 領域特定する手法の開発がすすんでいる．これらの手法は一般にマッピング，正規化，変化点の同定，領域の推定の 4 つの手順を踏む．まず，NGS より得られた各断片配列が，ヒト（対象種）のゲノムのどの位置に由来するかを検索，対応付ける (マップ)．対象の患者に CNV がない場合は参照ゲノム全体にほぼ均一にマップされるが (図 1)，逆に CNV がある場合には，その領域由来の断片数が異なるため，対応する領域に多数の，あるいは，少数の断片配列がマップされると考えられる (図 2)．

一般に，各領域にマップされる断片数には起伏があるため，平滑化のためゲノム配列を適当な長さの領域に区切り（以下ウインドウと呼ぶ)，各領域に対応しているリード数をカウントする．平均的なリード数に対し，重複が起きているウインドウは 1.5 倍，欠失が起きているウインドウでは 0.5 倍のリード数が対応すると考えられる．単一のウインドウでは信頼性が薄いと考えられるため，連続する複数のウインドウで重複，

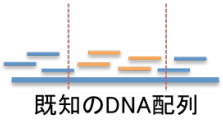

図 1: CNV なし

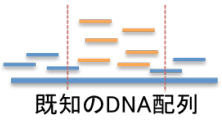

図 2: CNV あり

あるいは，欠失が推定された事をもって，CNV の存在を推定する [4]．

しかしこの手法には問題点がある．適切なウインドウ位置が設定できないと CNV が観測できない場合があるためである．CNV の境界がウインドウ位置と一致すればウインドウ間において急激なリード数の変化が観測され CNV を高精度に検出できるが，CNV の境界が，ウインドウの中央になった場合，ウインドウ内のリード数の変化がゆるやかになり，変化を推定することが困難となる．

この問題を避けるため CNV-seq[5] では，各データに対して適切なウインドウサイズを計算し，CNV 領域を検出している．

3　提案手法

本研究では移動平均を用いた手法を提案する．ノイズの無い環境ではコピー数の増加，減少を明確に区別可能だと考えられるが，実際観測されるリード数にはノイズが多く，コピー数の変化点が明確でないことが多い．そのようなデータからでも CNV を検出するために，金融分野や計測分野でよく用いられる移動平均の考え方を採用する．移動平均とは系列データを平滑化する手法である．

N 番目のウインドウ内のリード数を p_N とする．N 番目のウインドウを中心に M 個のウインドウを考えた時の平均リード数を

$$R(N,M) = \frac{p_{N-\lfloor \frac{|M|}{2} \rfloor} + p_{N+1} + ... + p_{N+\lceil \frac{|M|}{2} \rceil - 1}}{|M|}$$

とする．移動平均は，M を固定して，N を変化させることで作成される平均の値である．

$R(N,M)$ の定義と大数の法則により，$|M|$ が小さい時には N による $R(N,M)$ のバラつきが大きく，$|M|$ が大きい時には N による $R(N,M)$ のバラつきが小さいと考えられる．

ゲノム全体に関し，左から順に移動平均を求める場合を考えよう．図 3 から 5 で，リード数が青線のように変化している時，$|M|$ が狭い場合 (赤矢印) と $|M|$ が広い場合 (緑矢印) で平均値を比較する．図 3 のようなグラフであれば $|M|$ の大小によらず，平均値はほぼ等しい (黒丸)．しかし図 4 のような増加傾向にある場合，平均値は $|M|$ が大きい方 (緑丸) が低く，図 5 のような減少傾向にある場合，平均値は $|M|$ が大きい方

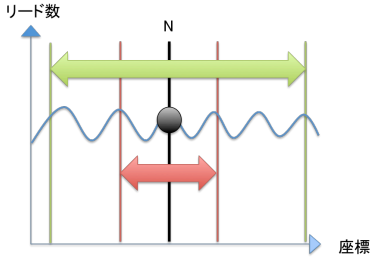

図 3: 通常

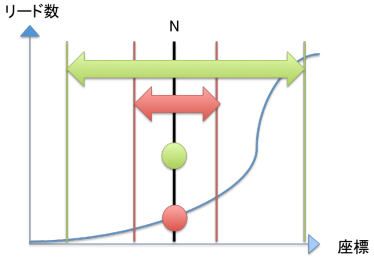

図 4: 増加傾向

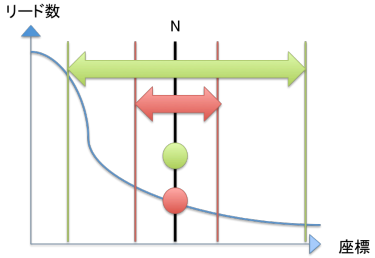

図 5: 減少傾向

(緑丸) が高くなる．大きな変化が起こった際に，$|M|$ が大きい時の方が緩やかに変化しているのがわかる．

急激に変化が起きる時，$|M|$ が大きいものと小さいものには値の差が生じるので，その差を乖離度と呼ぶ．乖離度がより大きくなる点を変化点とすることで，CNV をとらえる．狭い方を狭域移動平均，広い方を広域移動平均と呼ぶ．

また，ウインドウサイズは検出したい CNV のサイズに依存する．例えば最小でサイズ 1000 の CNV を検出したい場合，狭域移動平均のウインドウサイズを 1000 未満にする必要がある．

4　疑似データを用いた実験

疑似データはヒト (hg19) の 21 番染色体を基に重複を 20 カ所，欠失を 26 カ所挿入した疑似染色体配列を作成し，擬似的にリードを作成することで実験を行った．標準的な実験に合わせ，平均 30 リードが各塩基から読まれるように設定した．狭域移動平均領域サイズを (500,1000,2000)，広域移動平均領域サイズを (5000,10000,20000) でそれぞれ変動させたときに，狭域移動平均領域サイズを 1000，広域移動平均領域サイズを 10000 とした場合に最も精度が高かったため，今回はこの数値を採用した．

重複のリード数は基準値の 1.5 倍，欠失のリード数は基準値の 0.5 倍であるため，乖離度が基準値の半分以上あれば，その位置が変化点であることが分かる．しかし元データにはノイズが多く含まれているため，基準値の半分以下の乖離度でも，ある程度は変化点と見なすことが可能．今回は，乖離度が基準値の 0.4 倍以上である位置をリード数の変化点とした．結果として検出された範囲と，挿入した重複，欠失の位置を対応付け，正答率を表 1 にまとめた．CNV-seq でも同様に実験を行い，結果を表 2 にまとめた．感度=検出した CNV 数/全 CNV 数，特異度=検出した CNV 数/検出数である．

5　まとめ

今回提案した手法により，CNV 領域を検出することに成功した．従来法に比べ，特異度が高く偽陽性が少ないことが確認できたが，感度が低く，検出漏れが多かった．実用には更なる精度向上が必要である．

表 1: 本手法による結果

	○	×	計	感度	検出数	特異度
欠失	6	14	20	0.30	12	0.50
重複	12	14	26	0.46	13	0.92
計	18	28	46	0.39	25	0.72

表 2: CNV-seq による結果

	○	×	計	感度	検出数	特異度
欠失	20	0	20	1.00	-	-
重複	26	0	26	1.00	-	-
計	46	0	46	1.00	2083	0.03

謝辞

産業技術総合研究所　ゲノム情報研究センター
山形浩一 特別研究員に助言をいただきました

参考文献

[1] The 1000 Genomes Project Consortium.: An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56-65(2012).

[2] Welter D., *et al.*: The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 2014, Vol. 42, Database issue D1001-D1006(2014).

[3] The Cancer Genome Atlas Research Network.: The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* 45, 1113-1120(2013).

[4] Zhao M., *et al.*: Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics, 14(Suppl 11):S1(2013).

[5] Xie C. and Tammi MT.: CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics 10:80(2009).