2006－BIO－4（8）
2006／2／10

ヒトゲノム重複領域の網羅的解析

○ 佐藤慶治[1],坂手龍一[1,2],村上勝彦[1,2],武田淳一[1,2],松矢明宏[2,3],藤井康之[1,2],伊藤剛[1,4],五條堀孝[1,5],今西規[1]
所属 [1]産業技術総合研究所　生物情報解析研究センター 、[2]バイオ産業情報化コンソーシアム　生物情報解析研究センター、[3]日立製作所、[4]農業生物資源研究所　ゲノム研究グループ、[5]国立遺伝学研究所　生命情報・DDBJ研究センター

[要旨]
進化の過程におけるゲノム重複の役割を解明することを目的として、ヒトゲノム配列上の重複領域を全ゲノムレベルで網羅的に解析した。NCBI build35 の約 31 億塩基に及ぶヒトゲノム配列に対して BLASTZ を用いた自己アラインメントを作成し、重複領域を検出した。結果として、ヒトゲノム全体に占める重複領域の割合は 5.2%であり、そのうち約4割にあたる2.1%は同一染色体内での重複であった。これらの重複領域に対して重複の年代を推定し、重複が百万年あたり 4.0Mbp の速度で生成し、2300 万年の半減期にて減衰していることが判明した。

Comprehensive analysis of segmental duplications in the human genome

Yoshiharu Sato[1], Ryuichi Sakate[1,2], Katsuhiko Murakami[1,2], Jun-ichi Takeda[1,2], Akihiro Matsuya[2,3], Yasuyuki Fujii[1,2], Takeshi Itoh[1,4], Takashi Gojobori[1,5], Tadashi Imanishi[1]
[1]Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo , Japan,[2]Japan Biological Information Research Center, Japan Biological Informatics Consortium, Tokyo, Japan,[3]Hitachi Co., Ltd., Tokyo, Japan,[4]Genome Research Department, National Institute of Agrobiological Sciences, Tsukuba, Japan,[5]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Mishima, Japan

[Abstract]
We searched for the segmental duplications in the human genome sequences in order to assess the role of segmental duplication during evolution. A self-alignment of NCBI build 35 human genome (~3 billion bp) was performed using the genome alignment software BLASTZ. As a result, we found that 5.2% of the human genome contains segmental duplications, and ~40% (2.1% of the human genome) of them were intra-chromosomal. Moreover, we deduced the divergence time of segmental duplications and the duplicated genes embedded in them. We obtained estimates of 23 million years as the average half-life of duplicated regions, and 4.0 Mbp per million year as the rate of origin of duplicated regions.

[Introduction]
ゲノム重複(segmental duplication)はゲノムの進化において遺伝子の新規獲得やゲノムサイズの変化等に寄与しており、重複領域を全ゲノムレベルで解明することは、ヒトゲノムの進化の解明にとって非常に重要であると考

えられる。また、重複を引き起こすゲノム再編成は様々な遺伝病や悪性腫瘍に関連することが知られており[1,2,3]、遺伝子のマッピング情報や反復配列、転写制御領域といった各種のゲノムアノテーション情報と統合された重複領域データセットを整備することで、病気のメカニズムにせまる分野での研究推進に貢献することが期待される。

本研究では、相同性検索ツールである BLASTZ を用い、ヒトゲノム重複領域を全ゲノムレベルでその位置、構成を明らかにし、遺伝子情報との対応がついたデータセットを作成することを目的とする。また、そのデータを用い重複の年代推定を行い、その分布を求めることで、ヒトゲノムの重複の生成過程を明らかにし、遺伝子構造との対応をとることによってヒトゲノムの進化に関する知見を得ることを目指す。

[Methods]

配列の分割と相同性検索の実行

ヒトゲノムの配列として UCSC にて提供されている最新のアセンブル配列(hg17)をダウンロードし、反復配列が既に小文字でマスクされている配列を使用した。本研究では染色体上の位置が決定されていないランダムの配列とミトコンドリア配列は使用しなかった。相同性検索を行う入力配列として、各染色体の配列を 10Mbp ごとに 10Kbp の重なりを持たせて分割し、クエリ側の配列とした。サブジェクト側の配列は 50Mbp に 10Mbp の重なりにて分割し、相同性検索ツール BLASTZ を用いた総当たりの相同性検索を行った。ここでは、正確なアラインメントを取り、重複の端点を正確に同定するために、非反復配列部分を優先的にアラインメントし、伸張の際には反復配列を使用するという手法を用いた。そのために、BLASTZ の実行は"C=2,Y=3400,T=4"のオプションにて行った。各々分割された計算を 64CPU の PC クラスタで OpenPBS のシステムによって並列化処理したところ、全ての相同性検索にかかった実行時間はおよそ１日であった。

相同領域のフィルタリング

同一の領域が複数箇所に重複するような場合、実際に重複によって生成された相同部位間以外にも相同性の高い部位が存在するケースが考えられる。こういった相同部位を二重に重複部位としてカウントしてしまうのを防ぐ目的で、以下の方法にて相同部位のフィルタリングを行った。(i)全ての相同部位をスコアの高い順にソートする。(ii)スコアの高い相同部位からリストに入れていき、２つの領域を重複領域として記録する。このとき、２つの部位が両方ともすでに重複領域と同定されている場合、この相同部位は捨てる、という操作を全てのペアについて繰り返し、ベスト選定後の重複領域データセットとした。

他生物種との相同性検索の実行

重複領域の年代を推定する際にヒト・チンパンジー間の相同性の値を参照するため、上記と同様の手法(BLASTZ のオプションは C=2,Y=3400,T=4 を使用)を用いてヒト対チンパンジーの相同性検索を行った。ヒトゲノムは UCSC hg17、チンパンジーの配列は UCSC panTro1 の配列を使用した。

遺伝子のマッピング

RefSeq,Ensembl,H-InvDB,DDBJ に登録されている合計 242,650 の転写物を Blastn,Blat,Est2genome を用い hg17 のヒトゲノム配列にマッピングし、その位置、遺伝子構造を明らかにし、上記重複領域データとの対応を取った。

[Results & Discussion]

重複領域の分布

重複領域のヒトゲノム全体に占める割合、染色体ごとの分布、さらに染色体内、間での割合を表1に示した。ゲノム全体の5.2%が重複領域(1つ以上の重複領域を持つ領域)であり、染色体内の重複に限定しても2.1%の領域が重複領域であった。この比率から染色体間重複に比べ染色体内重複が非常に起こりやすいということが言える。さらに染色体ごとに結果をみると各々特徴が異なり、染色体内重複の占める割合についてはばらつきがあり、重複の生成の様式が染色体ごとに大きく異なるということが考えられる。性染色体や9番、16番染色体では重複領域の割合が高く、また3番、6番染色体では割合が低いという傾向がみられた。

表1 重複領域の染色体ごとの割合。括弧内は染色体内重複に限定した場合の数値をあらわす。

染色体	染色体の全長(bp)	重複領域(bp)	重複領域の割合(%)	染色体	染色体の全長(bp)	重複領域(bp)	重複領域の割合(%)
1	245,522,847	10,391,966	4.2 (2.0)	13	114,142,980	3,899,716	3.4 (0.9)
2	243,018,229	11,086,367	4.6 (2.2)	14	106,368,585	3,486,971	3.3 (0.9)
3	199,505,740	5,451,044	2.7 (0.9)	15	100,338,915	6,381,845	6.4 (3.1)
4	191,411,218	6,286,453	3.3 (1.1)	16	88,827,254	6,574,173	7.4 (4.0)
5	180,857,866	7,188,440	4.0 (2.1)	17	78,774,742	5,419,110	6.9 (3.6)
6	170,975,699	4,416,168	2.6 (0.9)	18	76,117,153	2,754,830	3.6 (2.9)
7	158,628,139	11,166,283	7.0 (3.9)	19	63,811,651	4,428,714	6.9 (2.6)
8	146,274,826	5,056,521	3.5 (1.3)	20	62,435,964	2,112,403	3.4 (0.8)
9	138,429,268	10,245,061	7.4 (4.1)	21	46,944,323	2,255,856	4.8 (0.3)
10	135,413,628	7,125,606	5.3 (2.5)	22	49,554,710	3,300,897	6.7 (1.7)
11	134,452,384	6,786,756	5.0 (2.8)	X	154,824,264	13,811,446	8.9 (2.1)
12	132,449,811	4,515,281	3.4 (0.9)	Y	57,701,691	14,445,964	25.0 (8.6)
				Total	3,076,781,887	158,587,871	5.2 (2.1)

重複領域の年代分布

重複領域の長さと配列一致度との関係を図1に示した。また、ヒト対チンパンジーのゲノムアラインメントによって得られた平均の配列一致度97.89%(gap込みの数値)という値、さらにヒトとチンパンジーがおよそ500万年前に分岐したという推定値を用いて各重複領域の年代推定を行った。するとその分布は生存曲線に非常に近い分布をとり、領域の長さにて計測した場合、殆どの重複領域が 0.1 以下の分岐度のもので占められていた。これは時間の経過によって変異が蓄積し、重複領域が徐々に減衰していくことを示している。次にLynch[4]らによって報告された手法を用いて重複領域の生成速度、消失速度を推定した。最小二乗法を用い、重複領域の長さの分布は相関係数R=0.94にて生存曲線に近似することができ(図2)、その近似式より、重複領域が百万年あたり4.0Mbpの速度にて生成し、2300万年の半減期にて減衰していくことが判明した。

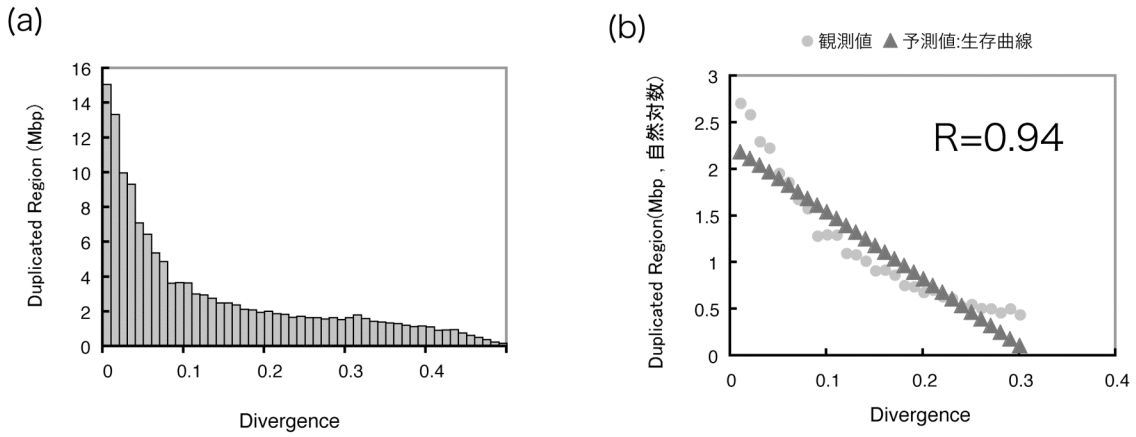

図 1 重複領域の合計長の分岐度に対する分布(a)と生存曲線への最小二乗法を用いた近似(b)

遺伝子情報との対応

ヒトゲノム上にて重複領域と RefSeq 遺伝子の位置情報の対応をとったところ、全 23,257 転写物のうち 455 の遺伝子が全長にわたり重複領域上に存在することが判明した。これらの遺伝子は重複を経て新規機能の獲得や偽遺伝子化、転写制御領域の進化、スプライシング変異体の多様化に関与している可能性が考えられる。その中で、選択的スプライシングのパターンが大きく異なっていたケースを図 2 に挙げた。この例では免疫関連遺伝子の遺伝子座が重複しているが、そのスプライシング変異体の組み合わせは 2 つの遺伝子座の間にて大きく異なり、さらに逆ストランド側の転写物のスプライシングの様式も異なっていた。さらに中央のカラムにゲノムアラインメントの対応が示されているが、変異の分布が一様ではないことがわかり、この情報を用い転写・翻訳制御領域に関しての重要な情報を得ることができると考えられる。この例から、重複がおこり、変異が蓄積し、転写制御やスプライシングの様式が多様化していくという進化のシナリオを推測することができる。

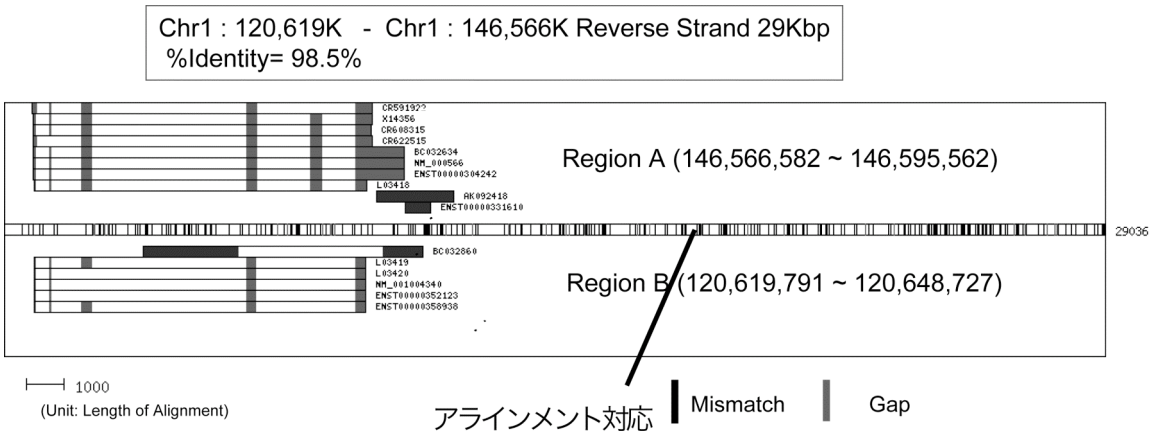

図 2 重複領域（ゲノムアラインメント）と遺伝子座の対応。色の薄い BOX はアラインメントと同方向のストランドに存在するエクソンを示しており、色の濃い BOX は逆ストランド側に存在するエクソンを表す。Region A 側の同方向側の遺伝子座は"Fc fragment of IgG"という機能アノテーションがつけられており、対応する Region B 側の遺伝子座では"Fc gamma receptor"であった。逆ストランド側で対応している 2 つの遺伝子座は機能未知であった。

[Conclusion]

全ヒトゲノムにおける重複領域を同定し、遺伝子情報と対応の取れた重複領域データセットを作成し、以下の知見を得た。(1) 重複領域はヒトゲノム全体のおよそ 5.2%を占め、染色体内での重複に限定すると 2.1%の領域が重複領域である。(2)重複領域は 4.0Mbp/Myr の速さにて生成しており、半減期は 23Myr である。また、ここで得られた重複領域のデータセットは重複のメカニズムの解明、さらに重複遺伝子、偽遺伝子、選択的スプライシング、転写制御領域などの研究に非常に有用であると考えられる。

[References]

[1] Shaw et al. (2004),*Human Molecular Genetics*, **13**, R57-R64

[2] Kolomietz et al. (2002),*Genes, Chromosomes & Cancer*, **35** , 97-112

[3] Emanuel et al. (2001),*Nature Reviews Genetics*, **2**, 791-800

[4] Lynch et al. (2000) *Science*,**290**,1151-1155