した[6]．このように後天性 LQTS の少なくとも一部には，通常は機能異常が顕性化しない遺伝子異常を持つものがあり，これらの変異キャリアは薬剤などの 2 次的要因を受けたときに初めて QT 延長が顕性化すると考えられる．

2．QT 短縮症候群（SQTS）

心筋の再分極は延長したときに不整脈を起こしやすくなるが，逆に極端に短縮したときも不整脈を来す．心電図 QT 間隔の短縮（QTc< 360 msec）を特徴とする遺伝性不整脈 SQTS に 3 つの原因遺伝子（*KCNH2*・*KCNQ1*・*KCNJ2*）が同定された[7]．これらの遺伝子はそれぞれ，LQT2・LQT1・LQT7 の原因遺伝子と同一であるが，SQTS 変異はKチャネル電流を増加し（gain－of－function）再分極を加速する点で LQTS のKチャネル変異と異なる．

3．Brugada 症候群

器質的心疾患がない心室細動（特発性心室細動）のうち，心電図右側胸部誘導の ST 上昇を特徴とする一群は Brugada 症候群と呼ばれる．Brugada 症候群は日本を含めた東アジアの罹患率が高いと言われる．現在同定されている唯一の遺伝子は *SCN5A* だが[8]，変異が同定されるのは患者全体の 20 ～ 30％ に過ぎない．また，日本人の 0.05 ～ 0.1％ には全く無症状でありながら Brugada 型の心電図を呈する症例（無症候性 Brugada 症候群）があるが，その病因は明らかではない．

4．カテコラミン誘発性多形性心室頻拍（CPVT）

CPVT は，運動によって誘発される多型性心室頻拍を特徴とし，青少年に好発する致死性遺伝性不整脈である．最近，CPVT 家系に心筋リアノジン受容体遺伝子 *RyR2* に変異が同定された．心筋の脱分極に伴って細胞膜のL型 Ca チャネルから Ca^{2-} が流入するが，これをトリガーとして筋小胞体（SR）から大量の Ca^{2+} が放出される．リアノジン受容体は SR からの Ca^{2+} 放出を行うイオンチャネルで，心筋の興奮収縮連関（EC coupling）において中心的な役割を果たす．現在までに報告された *RyR2* 変異は 30 種類以上にものぼるが，その多くは 5000 アミノ酸という巨大分子のごく限られた部分に集中している[9]．変異リアノジン受容体は，アドレナリン刺激よる SR からの

表1　心筋症の新分類（文献[15]より引用）

1．原発性心筋症（primary cardiomyopathy）

a．遺伝性（genetic）
肥大型心筋症（HCM），不整脈源性右室心筋症（ARVC），左室緻密化障害（LVNC），グリコーゲン蓄積症（PPKAG2，Danon病），心臓伝導障害，ミトコンドリア心筋症，心筋イオンチャネル病（LQTS，SQTS，Brugada症候群，CPVT，SUNDS）

b．混合性（mixed）
拡張型心筋症（DCM），拘束型心筋症（RCM）

c．後天性（acquired）
炎症性（心筋炎）心筋症，ストレス誘発性（タコツボ型心筋症），心外膜炎，頻拍誘発性心筋症，インスリン依存性DM新生児心筋症

2．二次性心筋症（secondary cardiomyopathy）

略語：巻末の「今号の略語」参照

Ca放出を増強し，本来細胞内Ca濃度が低い拡張期においても細胞内にCaを漏出するため，遅延後脱分極（DAD）が発生し，心室性不整脈を来す．

5．心房細動（AF）

AFは最も頻度の高い不整脈の1つで，そのうち明らかな原因のないlone AFは15～30％を占める[10]．AF全体の5％に家族歴があり[11]，遺伝的な要因も関与していると思われる．最近，家族性AFの家系にKチャネル遺伝子*KCNQ1*・*KCNE2*・*KCNJ2*のgain-of-function変異が同定された[12~14]．これらの変異は活動電位持続時間を短縮するため，心房の有効不応期が短縮し，リエントリーが生じやすくなり，AFが出現すると考えられる．一方，AFの多くは明白な家族内発生はみられず，後天的な環境要因と幾つかの遺伝要因の組み合わせで発症すると考えられるが，いわゆるAF感受性遺伝子は明らかになっていない．

6．心筋症

近年の分子遺伝学の急速な発達によって，心筋症の原因遺伝子が次々と解明され[1]，新しい疾患概念も生まれた．それとともに1995年WHOの定義にも改定が必要となり，最近，心筋症に関する新たな定義と分類が提唱された（表1）[15]．「心筋症は，心筋の機械的・電気的機能不全を伴うさまざまな心疾患で，多くは心臓の肥大や拡張を伴い，遺伝要因を含めたさまざまな病因を有する．異常が心臓に限局

していることもあるが，全身疾患の心臓病変であることもあり，多くは心臓死や進行性の心不全の原因となる」と定義されている．心筋症は，病変が心臓に限局しているか，全身疾患の心病変であるかによって，原発性心筋症（primary cardiomyopathy）と二次性心筋症（secondary cardiomyopathy）に大きく二分され，原発性心筋症はさらに遺伝性（genetic），混合性（mixed），後天性（acquired）に分類される．遺伝性心筋症には肥大型心筋症（HCM），グリコーゲン蓄積症，不整脈源性右室心筋症（ARVC），左室緻密化障害（LVNC），心臓伝導障害，ミトコンドリア心筋症，心筋イオンチャネル病が含まれる．HCMにはβミオシン重鎖を始めとする多くの原因遺伝子が報告されている．AMP activated protein kinase γ-2サブユニット（PPKAG2）や，lysosome-associated membrane protein 2（LAMP-2）の遺伝子異常によるDanon病[16]は，グリコーゲン蓄積によって心肥大と早期興奮を来す．ARVCには9つの遺伝子座（ARVC1～9）が報告されおり，そのうちCPVTの原因遺伝子でもある*RyR2*を含め，desmoplakin，plakophillin-2，TGF-β3[17]，desmoglein-2[18]と5つの原因遺伝子が明らかになっている．特筆すべきなのは，LQTSやBrugada症候群のような心筋イオンチャネル病が新たに原発性心筋症に分類されたことである．これは，イオンチャネルの遺伝子異常は単に活動電位の異常を引き起すだけでなく，イオンチャネルと連関するさまざまなタンパクの構造や機能にも影響を及ぼすという仮説に基づいている．一方，混合性心筋症には，拡張型心筋症（DCM）と拘束型心筋症（RCM）が分類され，後天性心筋症には炎症性（心筋炎），ストレス誘発性（タコツボ型心筋症），心外膜炎，頻拍誘発性心筋症などが分類されている．二次性心筋症にはアミロイドーシス，サルコイドーシスなどを始めとする数多くの心筋症が分類されている．

多因子遺伝子疾患

心血管疾患，がん，糖尿病のような頻度の高い疾患は，複数の遺伝要因と環境要因の組み合わせによって起る．これらの疾患に何らかの遺伝性がみられるとき，共通のvariantがその原因となっていると推測される（common disease common variant hypothesis）．ポストゲ

ノム時代の遺伝子学に期待されるのは，これらの多因子遺伝子疾患の遺伝要因を特定し，遺伝子情報に基づいた個別化医療を開始するための基盤を確立することであり，循環器領域においても，高血圧・心筋梗塞などの疾患感受性遺伝子に関する研究が進んでいる[1][19~21].

多因子遺伝子疾患に対する研究アプローチは，多発家系を用いた連鎖解析，関連解析（association analysis），同胞対解析（sib－pair analysis）などが用いられる．関連解析は，患者（ケース）群と対象（コントロール）群の間で一塩基多型（SNP）などの遺伝子多型に出現頻度の差があるかを解析する（case－control association study）．一般に，2つの任意の遺伝子マーカーは，ゲノム上での距離が十分に近ければ，その間に減数分裂時の染色体組換えが入る確率が低くなるため，連鎖が保たれる（連鎖不均衡 LD）．したがって，患者・対照の両群間で SNP の出現頻度に差があった場合，その SNP が遺伝子の量的調節やタンパク機能を変化させる可能性とともに，その SNP が疾患関連遺伝子の近傍マーカーとなっている可能性も考えられる．もし，組換えが染色体で均一かつランダムに起るとすれば，疾患にかかわる遺伝要因を網羅的に検索するには，多数の患者と対象群ですべての SNP をタイピングするという非現実的な作業が必要となる．しかし，組換えは染色体上で均一に起きているのではなく，ゲノムは，高頻度に起る部位（組換えホットスポット）と連鎖不均衡の強い部分（LD ブロック）という，分節構造をとることが最近の研究から明らかになった[22]．したがって，組換えがほとんど起らない LD ブロック内に存在する SNP の組み合わせ（ハプロタイプ）を慎重に選んで遺伝子型を判定すれば，この領域で高頻度に存在するハプロタイプを特定するのに必要な「タグ」SNP はわずかな数ですむ．

多因子遺伝子疾患の疾患感受性遺伝子をゲノムワイドで効率良く探索するためには，ゲノムの LD 構造と代表的な SNP タイピングに関する体系的・網羅的なデータが必要である．この目的のために，国際ハップマッププロジェクトが 2002 年に発足し，アフリカ，アジアとヨーロッパを起源とする 269 人の DNA サンプルを用いて 100 万ヵ所以上の一般的な SNP のタイピングが行われ，ハップマップ（HapMap：ハプロタイプ地図の略称）が作成された．その結果は昨

年，ヒトゲノムに存在する頻度の高い一般的な多型の公共データベースとして最初のデータが公開された[23]．

1．心筋再分極を規定する遺伝子素因

LQTS や SQTS のような遺伝子変異は，心筋再分極を極端に延長・短縮する，不整脈器質の遺伝的な“決定因子”であるが，心電図 QT 時間は正常人でもばらつきがあり，心拍数，年齢，性，電解質，薬物などの要因によっても影響を受ける．最近，心筋再分極を規定する遺伝子素因に関して，候補遺伝子アプローチの研究と HapMap を利用したゲノムワイド関連研究が相次いで報告された．

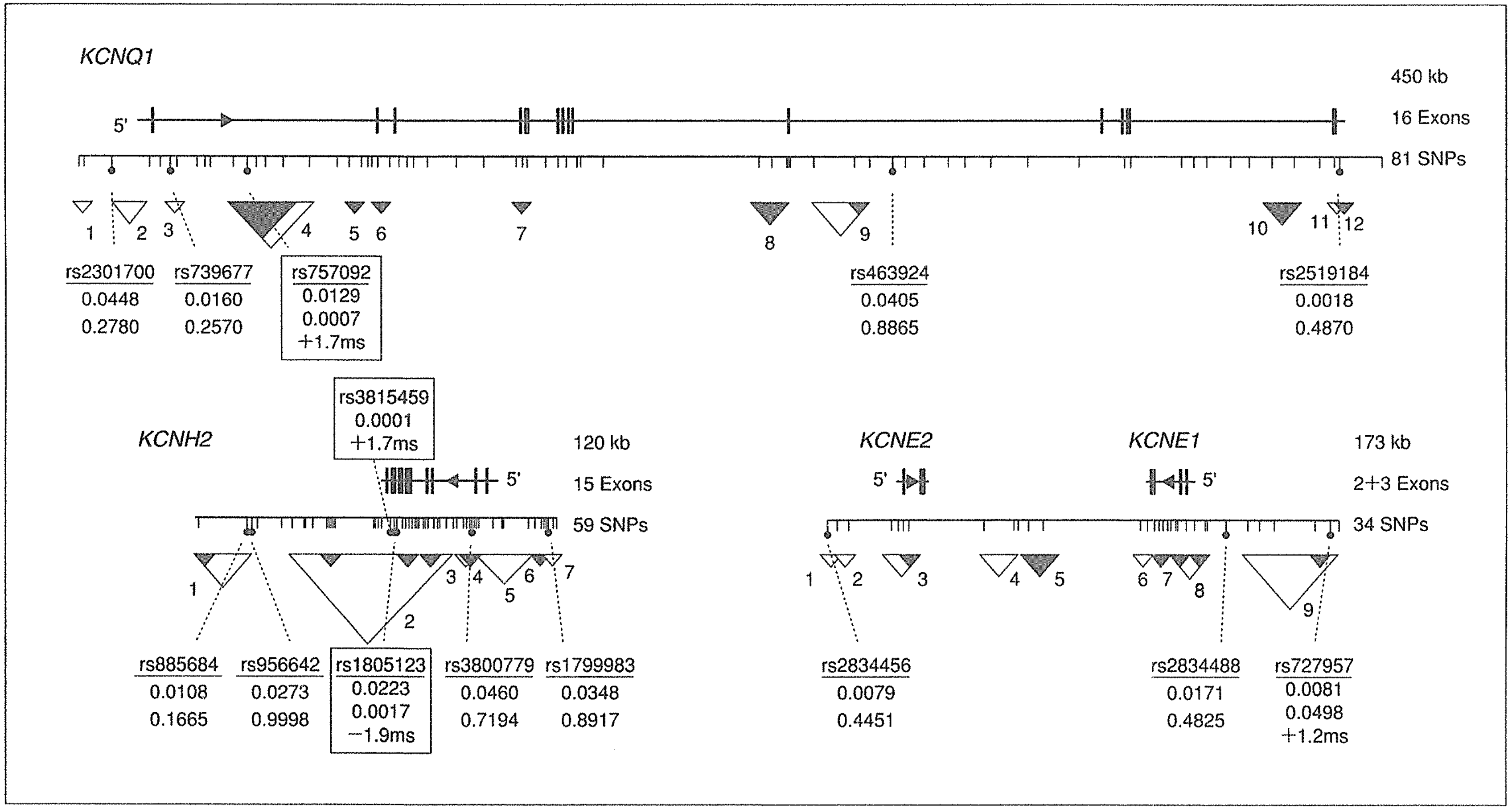
1）候補遺伝子アプローチ

LQTS 遺伝子の多型は正常人にも見られるが，その一部は再分極に影響を与え，時に薬剤誘発性 LQTS の原因となることがある[4]．再分極の感受性因子を明らかにする1つの方法論は，LQTS／SQTS の8つの原因遺伝子に注目し，その遺伝子多型を患者群と対照群で比較する集団関連研究（population based association study）である．Pfeufer らは，正常人 689 人（KORA study）のゲノムを用いて，先天性 LQTS の原因遺伝子である4つのKチャネル（*KCNQ1*, *KCNH2*, *KCNE1*, *KCNE2*）上の 174 個の SNP のタイピングを行い，QT_RAS 値（QT 時間を性・年齢・脈拍数で補正した値）との関連を調べた（図1）[24]．QT と有意な関連がみられた 14 個の SNP について，さらに 3,277 人のゲノムで2次スクリーニングしたところ，*KCNQ1* のイントロン1にある SNP が QT を平均 1.7 ms 延長させることが判明した．また，*KCNH2* の LD ブロック2には，QT を 1.9 ms 短縮するアミノ酸置換（K897T）型の SNP[25] と，1.7 ms 延長する SNP を同定した．この3つの SNP に関して QT 延長させる対立遺伝子をホモで持つ集団と，全く持たない集団の QT 時間を比較すると，平均 10.5 ms の違いがあった．この事実は，これらのKチャネル遺伝子の SNP は単独では QT 時間に与える影響は小さいが，それらが組み合わさることによって効果が相加的に増強する「心筋再分極の遺伝的修飾因子」であることを示す．

2）ゲノムワイド集団相関解析

心筋の再分極は LQTS 原因遺伝子の変異や多型ばかりではなく，

図1 QT 延長症候群（LQTS）関連Kチャネル遺伝子の構造（連鎖不均衡とSNP）（文献[21]より引用改変）



4つの LQTS 関連Kチャネル遺伝子について，最上段にエクソン（縦線）と遺伝子の方向（矢印）が，その下に，1次スクリーニングに用いた 174 個の SNP（｜）と，2 次スクリーニングに用いたSNP（●）の位置が示されている．それぞれの SNP には dbSNP の ID 番号，p値，マイナー対立遺伝子1個による平均 QT_RAS 時間の変化が記されている．また，連鎖不均衡（LD）を示す遺伝子領域（ブロック）は尺度（D'：▽）と r^2（>0.5, ▼）で表示され，番号がつけられている．*KCNQ1* イントロンの rs757092 と，*KCNH2* の LD ブロック2上の SNP（rs1805123, rs3815459）は本研究で明らかになった再分極感受性 SNP である．

環境要因やその他の遺伝要因によって規定されている．心筋再分極を修飾する遺伝要因を特定するために，Arking らは KORA S4 コホート 3,966 人の心電図とゲノム DNA を用いた集団関連研究を行った[26]．第1段階は，全コホートのうち QT－RAS 時間の最も長い女性 100 人と短い女性 100 人に対して，115,000 個の SNP を用いた全ゲノムタイピングを行い，QT－RAS 時間と関連の強い 10 個の SNP を同定した．また，心筋再分極に関与する 45 個の候補遺伝子についても SNP タイピングを行い，関連のある 10 個の SNP を同定した．1次スクリーニングで陽性になった SNP について，両群とも 200 人ずつ増やした2次スクリーニングによって陽性遺伝子を8個に絞り込み，最終的に残り全員による3次スクリーニングで，再分極の修飾遺伝子として *NOS1AP*（CAPON）を同定した．さらに，別のコホート KORA F3，Framingum Heart Study のゲノム解析でも，この結果の正当性が確認された．*NOS1AP* は，神経組織で強く発現する NOS1（nNOS）の修飾タンパク CAPON の遺伝子である．CAPON は NOS1 のC末端にある PDZ ドメインに結合するタンパクで，NMDA 受容体との共役を修飾することが知られているが[27]，心臓における機能は十分に解明されていない．しかし，心臓では NOS1 は筋小胞体に発現しているので，*NOS1AP* は Ca 遊離を介して心筋の再分極過程に影響を与える可能性もある．今後，国際 HapMap プロジェクトの推進によってさらに高密度の SNP マップが完成すれば，心筋再分極を規定する量的形質遺伝子座（QTL）を明らかにすることができると思われる．

また最近，Bezzina らは *SCN5A* のプロモータ領域にある6個の SNP が連鎖不均衡を示すことを認めた[28]．この LD ブロックの SNP の遺伝子型（ハプロタイプ）はほぼ2種類（HapA，HapB）に特定される．HapB は白人・黒人にはなく日本人の約 25％ に見られる．また，HapB は *SCN5A* プロモータの転写活性を低下させ，心室内伝導遅延と相関する遺伝型であることから，日本人を含める東アジアで罹患率が高いとされる Brugada 症候群に HapB が関与している可能性がある．

2．薬理ゲノミクス

薬理ゲノミクスは，薬剤反応性に影響を与える遺伝子情報（遺伝子多型）を探索することによって，より有効で安全な治療の確立を目指す学問である．循環器領域においても，HMG－coA 還元酵素[29]，apo E[30] などの遺伝子多型が，スタチンの有効性や副作用発現を予測しうる有効な遺伝子多型であることが判明した．また，βアドレナリン受容体の多型は心不全のリスクに連関すると同時に，心不全におけるβ遮断薬療法の有効性にも連関している．最近，β2 受容体の多型 N27G が突然死のリスクであるという報告もされた[31]．また，抗凝固薬ワーファリンの至適投与量は CYP2C9[32] とビタミンKエポキシド還元酵素[33] の遺伝子多型によって大きく影響される．

おわりに

循環器疾患の遺伝子異常に関して，一部の単一遺伝子疾患と，多因子遺伝子疾患の最近の進歩について概説した．単一遺伝子疾患の原因遺伝子とその機能異常の研究から，我々は病気の成り立ちについて遺伝子レベル・分子レベルでの理解が急速に進んだ．しかし，単一遺伝子疾患において同じ変異を持ったものが同じ臨床像を示すとは限らないし，逆に異なる遺伝子異常が同一の臨床像を示すこともある．このような遺伝型と表現型の乖離には，環境要因や何らかの修飾遺伝子やゲノムの多様性が関与していると推測される．今後，この分野の研究がますます進み，循環器領域の個別化医療が実現することを願う．

文献

1）木村彰方: 循環器系疾患の遺伝子学．最新医 59: 2017–2038, 2004.

2）Splawski I, et al: Cav1.2 calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism. Cell 119 (1): 19–31, 2004.

3）Napolitano C, et al: Evidence for a cardiac ion channel mutation underlying drug–induced QT prolongation and life–threatening arrhythmias. J Cardiovasc Electrophysiol 11 (6): 691–696, 2000.

4）Sesti F, et al: A common polymorphism associated with antibiotic–induced cardiac arrhythmia. Proc Natl Acad Sci USA 97 (19): 10613–10618, 2000.

5）Splawski I, et al: Variant of SCN5A sodium channel implicated in risk of cardiac arrhythmia. Science 297 (5585): 1333–1336, 2002.

6）Makita N, et al: Drug–induced long–QT syndrome associated with a subclinical SCN5A mutation. Circulation 106 (10): 1269–1274, 2002.

7) Brugada R, et al: Sudden death associated with short-QT syndrome linked to mutations in HERG. Circulation 109 (1): 30–35, 2004.

8) Chen Q, et al: Genetic basis and molecular mechanism for idiopathic ventricular fibrillation. Nature 392 (6673): 293–296, 1998.

9) Priori S G, et al: Cardiac and skeletal muscle disorders caused by mutations in the intracellular Ca^{2+} release channels. J Clin Invest 115 (8): 2033–2038, 2005.

10) Wiesfeld A C, et al: Genetic aspects of atrial fibrillation. Cardiovasc Res 67 (3): 414–418, 2005.

11) Darbar D, et al: Familial atrial fibrillation is a genetically heterogeneous disorder. J Am Coll Cardiol 41 (12): 2185–2192, 2003.

12) Chen Y-H, et al: KCNQ1 gain-of-function mutation in familial atrial fibrillation. Science 299 (5604): 251–254, 2003.

13) Yang Y, et al: Identification of a KCNE2 gain-of-function mutation in patients with familial atrial fibrillation. Am J Hum Genet 75 (5): 899–905, 2004.

14) Xia M, et al: A Kir2.1 gain-of-function mutation underlies familial atrial fibrillation. Biochemical and Biophysical Research Communications 332 (4): 1012–1019, 2005.

15) Maron B J, et al: Contemporary definitions and classification of the cardiomyopathies: an American Heart Association scientific statement from the council on clinical cardiology, heart failure and transplantation committee; quality of care and outcomes research and functional genomics and translational biology interdisciplinary working groups; and council on epidemiology and prevention. Circulation 113 (14): 1807–1816, 2006.

16) Yang Z, et al: Danon disease as an underrecognized cause of hypertrophic cardiomyopathy in children. Circulation 112 (11): 1612–1617, 2005.

17) Beffagna G, et al: Regulatory mutations in transforming growth factor-b3 gene cause arrhythmogenic right ventricular cardiomyopathy type 1. Cardiovascular Research 65 (2): 366–373, 2005.

18) Pilichou K, et al: Mutations in desmoglein-2 gene are associated with arrhythmogenic right ventricular cardiomyopathy. Circulation 113 (9): 1171–1179, 2006.

19) Ozaki K, et al: Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet 32 (4): 650–654, 2002.

20) 名倉 潤: 多因子遺伝病としての循環器疾患－高血圧－. 最新医 60: 2042–2049, 2005.

21) 尾崎浩一, 他: 多因子遺伝病としての循環器疾患－心筋梗塞－. 最新医 60: 2027–2034, 2005.

22) Daly M J, et al: High-resolution haplotype structure in the human genome. Nat Genet 29 (2): 229–232, 2001.

23) The International HapMap C: A haplotype map of the human genome. Nature 437 (7063): 1299–1320, 2005.

24) Pfeufer A, et al: Common variants in myocardial ion channel genes modify the QT interval in the general population: results from the KORA study. Circ Res 96 (6): 693–701, 2005.

25) Bezzina C R, et al: A common polymorphism in KCNH2 (HERG) hastens cardiac repolarization. Cardiovasc Res 59 (1): 27–36, 2003.

26) Arking D E, et al: A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. Nat Genet advanced online publication, 2006.

27) Jaffrey S R, et al: CAPON: A protein associated with neuronal nitric oxide synthase that regulates its interactions with PSD95. Neuron 20 (1): 115–124, 1998.

28) Bezzina C R, et al: Common sodium channel promoter haplotype in Asian subjects underlies variability in cardiac conduction. Circulation 113 (3): 338–344, 2006.

29) Chasman D I, et al: Pharmacogenetic study of statin therapy and cholesterol reduction. JAMA 291 (23): 2821–2827, 2004.

30) Gerdes L U, et al: The apolipoprotein e4 allele determines prognosis and the effect on prognosis of simvastatin in survivors of myocardial infarction : a substudy of the Scandinavian Simvastatin Survival Study. Circulation 101 (12): 1366–1371, 2000.

31) Sotoodehnia N, et al: Beta2–adrenergic receptor genetic variants and risk of sudden cardiac death. Circulation 113 (15): 1842–1848, 2006.

32) Higashi M K, et al: Association between CYP2C9 genetic variants and anticoagulation–related outcomes during warfarin therapy. JAMA 287 (13): 1690–1698, 2002.

33) Rieder M J, et al: Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. N Engl J Med 352 (22): 2285–2293, 2005.

BMC Genomics

BioMed Central

Research article

Open Access

Linkage disequilibrium of evolutionarily conserved regions in the human genome

Mamoru Kato[1], Akihiro Sekine[1], Yozo Ohnishi[1], Todd A Johnson[1], Toshihiro Tanaka[1], Yusuke Nakamura[1,2] and Tatsuhiko Tsunoda*[1]

Address: [1]SNP Research Center, RIKEN, Yokohama, Japan and [2]Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo, Japan

Email: Mamoru Kato - kato@src.riken.jp; Akihiro Sekine - sekine@genome.med.kyoto-u.ac.jp; Yozo Ohnishi - ohnishi@ims.u-tokyo.ac.jp; Todd A Johnson - tjohnson@src.riken.jp; Toshihiro Tanaka - toshitan@ims.u-tokyo.ac.jp; Yusuke Nakamura - yusuke@ims.u-tokyo.ac.jp; Tatsuhiko Tsunoda* - tsunoda@src.riken.jp

* Corresponding author

Published: 28 December 2006

BMC Genomics 2006, 7:326 doi:10.1186/1471-2164-7-326

Received: 24 May 2006

Accepted: 28 December 2006

This article is available from: http://www.biomedcentral.com/1471-2164/7/326

Abstract

Background: The strong linkage disequilibrium (LD) recently found in genic or exonic regions of the human genome demonstrated that LD can be increased by evolutionary mechanisms that select for functionally important loci. This suggests that LD might be stronger in regions conserved among species than in non-conserved regions, since regions exposed to natural selection tend to be conserved. To assess this hypothesis, we used genome-wide polymorphism data from the HapMap project and investigated LD within DNA sequences conserved between the human and mouse genomes.

Results: Unexpectedly, we observed that LD was significantly weaker in conserved regions than in non-conserved regions. To investigate why, we examined sequence features that may distort the relationship between LD and conserved regions. We found that interspersed repeats, and not other sequence features, were associated with the weak LD tendency in conserved regions. To appropriately understand the relationship between LD and conserved regions, we removed the effect of repetitive elements and found that the high degree of sequence conservation was strongly associated with strong LD in coding regions but not with that in non-coding regions.

Conclusion: Our work demonstrates that the degree of sequence conservation does not simply increase LD as predicted by the hypothesis. Rather, it implies that purifying selection changes the polymorphic patterns of coding sequences but has little influence on the patterns of functional units such as regulatory elements present in non-coding regions, since the former are generally restricted by the constraint of maintaining a functional protein product across multiple exons while the latter may exist more as individually isolated units.

Background

Linkage disequilibrium (LD) is non-random association between alleles at different loci and helps us to reconstruct the genetic history of human populations and to improve our understanding of the biological processes of recombination and natural selection [1]. LD also helps association studies to identify haplotypes that are linked to disease-causing variations. Early studies of LD focused on small sets of genes, such as the HLA genes [2] or the growth hormone gene cluster [3]. Recently, large-scale genotyping studies [1,4-6] have investigated the genomic patterns of LD in the human genome and have found considerable variation in its values, even for SNP pairs that are separated by identical physical distances. Some studies have tried to associate this variation with sequence features existing in the genome and found that genic or exonic regions are associated with strong LD in human populations. For example, extended LD regions are significantly overpopulated with SNPs located in genic or coding regions [5], and LD is stronger between exonic variants within a gene compared with intronic or intergenic SNPs [6]. The recent International HapMap Project also revealed an excess number of genes with strong LD on a genomic scale [7]. These findings can be explained by the previous hypothesis that purifying selection leads to an increase of LD [8]. This basic hypothesis in turn suggests that LD might be stronger in regions conserved among species than in non-conserved regions, since regions exposed to purifying selection tend to be conserved over evolutionary time.

In the present study, using the complete human/mouse sequences and polymorphic data from the HapMap Project, we unexpectedly observed that LD was significantly weaker in conserved regions than in non-conserved regions. A similar tendency was recently reported in a companion paper of the HapMap project [9]. To investigate this inconsistency between the predicted and observed tendencies, we examined the possibility that the relationship between LD and conserved regions is distorted by other sequence features, such as physical distance, genic content, GC/CpG frequency, or chromosomal location. However, these features were independent of the weaker LD tendency in conserved regions. Finally, we found that LD patterns carried by interspersed repeats were associated with this discrepancy. To precisely understand the relationship between LD and sequence conservation, we removed the effect of repetitive elements from the LD patterns, and found that although the previous hypothesis is partly correct, the reality is more complex than expected. That is, sequence conservation itself is not well associated with the degree of LD; however, on conserved coding regions only, it is related to a strong degree of LD. The results of our detailed analysis of the LD tendency in conserved regions imply that selective force produces a more complicated tendency in polymorphic patterns that varies according to the long-range or short-range functionality of DNA sequences.

Results

LD within conserved regions

We calculated pairwise r^2 and $|D'|$ values within conserved and non-conserved regions across the human genome and found that conserved regions contained lower proportions of SNP pairs that were in complete or nearly complete LD ($r^2 > 0.8$, $|D'| > 0.9$) when calculated as a function of physical distance (Fig. 1A for CEU and Additional file 1 for CHB, JPT, and YRI). A permutation test confirmed the significance of this observation ($p < 10^{-4}$ for all 10 kb bins of distance up to 40 kb; see Methods). We confirmed that allele frequencies had no effect on this result (data not shown). Since all results described here and below had the same tendencies for both r^2 and $|D'|$, we show only the r^2 results. We further checked the result by fine-scale recombination rates from the HapMap data [7] and found a higher recombination rate (1.41 cM/Mb on average) in conserved regions than that (1.26 cM/Mb) in non-conserved regions. This result is consistent with the LD results, since, in general, lower LD values are widely known to be related to higher recombination rates [8,10,11].

The finding of lower LD in conserved regions is inconsistent with the hypothesis that purifying selection increases the extent of LD [8]. Therefore, we first considered the possibility that the unexpected decrease of LD in conserved regions (i.e., the increase of LD in non-conserved regions) was distorted by the presence of genes, since genic regions had previously been shown to exhibit increased LD [5]. For that purpose, we took intersected regions of conserved/non-conserved regions with genic/non-genic regions, and generated datasets for four classes of regions: conserved genic, conserved non-genic, non-conserved genic, and non-conserved non-genic. Figure 1B (and Additional file 1) shows that conserved regions still possessed lower proportions of SNP pairs in strong LD compared to non-conserved regions for both genic and non-genic classes. Thus, gene content does not account for this effect. Next, since centromeric regions show stronger LD than telomeric regions [6,12], we also checked the possible involvement of chromosomal location by intersecting conserved and non-conserved regions with telomeric, centromeric, and other residual regions (see Methods). Figure 1C (and Additional file 1) shows the same tendency for all centromeric, telomeric, and residual regions. This result suggested that the weak LD in conserved regions was independent of chromosomal location.

In view of these results, we considered whether other factors, such as GC-content or CpG dinucleotides, may have been involved in the weak LD in conserved regions, because it was recently found that GC-content is associated with weak LD on a genomic scale [7,9]. However, GC-content and CpG dinucleotides are unlikely to account for the observed LD differences, since the proportions of their bases in conserved regions were almost equal to those in non-conserved ones (Additional file 2).

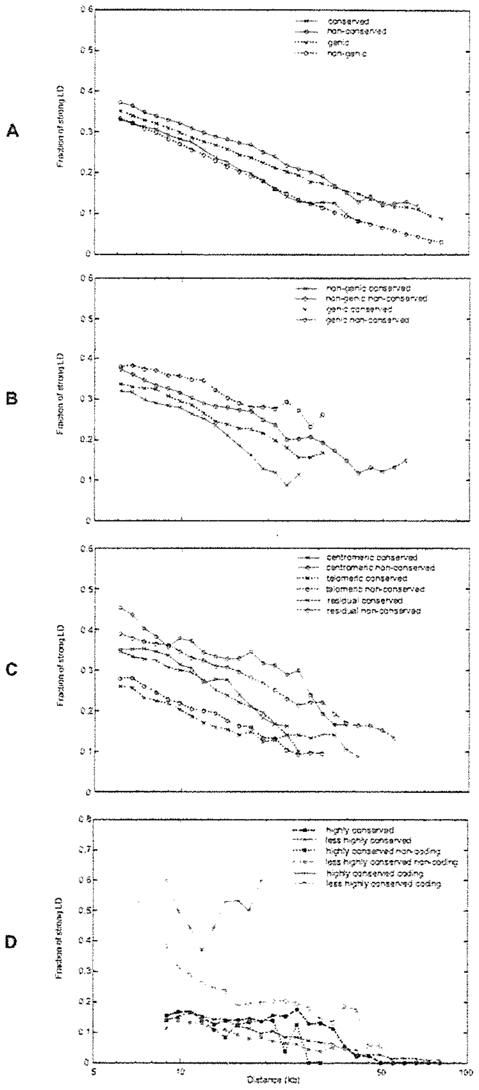


Figure 1

A moving average of the fraction of complete or nearly complete LD ($r^2 > 0.8$) versus distance between SNPs. All panels are those for CEU. See Additional file 1 for CHB, JPT, and YRI, which show the same tendency. (A) Plots of LD within DNA sequences conserved between the human and mouse genomes (in red with Xs), non-conserved regions (regions other than conserved ones; shown in red with circles), genic regions (in blue with Xs), and non-genic regions (in blue with circles). (B) Plots of LD within intersections of non-genic regions with conserved (in red with Xs) and non-conserved (in red with circles) regions, and of genic regions with conserved (in blue with Xs) and non-conserved (in blue with circles) regions. (C) Plots of LD within intersected regions of centromeric regions (the 10% definition, we only show plots in the 10% definition because of the same tendency in the 5% definition) with conserved (in red with Xs) and non-conserved (in red with circles) regions, of telomeric regions with conserved (in blue with Xs) and non-conserved (in blue with circles) regions, and of the residual regions (neither centromeric nor telomeric) with conserved (in green with Xs) and non-conserved (in green with circles) regions. (D) LD fractions for SNP pairs within highly conserved and less highly conserved regions (black and green), highly and less highly conserved non-coding regions (blue and light blue), and regions enriched (>20% in the bases) with highly and less highly conserved coding regions (red and pink). We selected only regions where the proportion of repeats was <20%, and since after this adjustment we found outliers of LD related to extreme GC-content, we further selected regions where the GC-content was 45–65%.

To verify this, we executed permutation tests (see Methods), and found that even when the effect of these sequence features was subtracted from LD, the adjusted LD in conserved regions was still significantly weaker than that in non-conserved ones ($p < 10^{-4}$).

The influence of repetitive elements

Next, we considered whether the weak LD tendency in conserved regions might be related to a lack of interspersed repeats in these regions, since interspersed repeats were recently reported to be related to strong LD on a genomic scale [7,9]. We found that the proportion of the total number of bases in repeats within conserved regions was half of the proportion found within non-conserved ones (Additional file 2), as previously observed [13]; this was probably because local rates of neutral variation may be low in conserved regions [13] or because selective pressure working around conserved regions may have excluded repetitive elements that would cause deleterious changes in the genome, such as changes in a gene's structure [14]. Indeed, our permutation tests showed that, after subtracting the effect of repeats from LD by regression, we no longer observed any significant difference in LD between conserved and non-conserved regions (p = 0.522). We confirmed these results by partial correlation analysis (Additional file 3). These findings suggest that the lack of repetitive elements accounts for weak LD in conserved regions. Among the several types of repeats, LINE/L1s had the largest regression coefficient in the regression analysis between LD and the proportion of bases contained in repeats (Additional file 2), the smallest proportion of bases in conserved regions compared to non-conserved ones (Additional file 2), and the largest total number of bases in the human genome (Additional file 2). Therefore, L1s appeared to mostly account for the weak LD in conserved regions.

Since we found that repetitive elements are strongly associated with weaker LD in conserved regions, we adjusted for the base-pair proportion of repeats as well as GC-content; the latter due to outliers of LD related to extreme GC-content after the repeat adjustment. We then compared the LD levels in highly conserved regions with those in less highly conserved regions. We expected that highly conserved regions would have stronger LD because the selective pressure on these regions was considered to be stronger. However, unexpectedly, we found no enhancement of the strong LD fraction within highly conserved regions compared to less highly conserved regions (Fig. 1D and Additional file 1). We then classified these regions into two groups, those enriched with coding sequences and those enriched with non-coding sequences. As a result, we found that regions enriched with highly conserved coding sequences had stronger LD than regions enriched with less highly conserved coding sequences. Meanwhile, no difference in LD was found between highly conserved non-coding regions and less highly conserved non-coding regions. To further confirm these results, we used fine-scale recombination rates from the HapMap data [7] and calculated average recombination rates for the same regions (see Methods). This method revealed a similar tendency, with a somewhat smaller recombination rate shown in highly conserved non-coding regions than in less highly conserved non-coding regions, while the difference was far greater between highly and less highly conserved coding rich-regions (Table 1). These results suggest that purifying selection that works on evolutionarily conserved regions surely increases the LD level in a series of coding sequences; however, it does not do so in non-coding sequences, as discussed below.

Discussion

Throughout the evolutionary history of a population, a variety of factors influence the LD level, such as recombination, mutation, genetic drift, natural selection, and demographic events [7,8,11,15]. Among these factors, natural selection is considered to generally increase the degree of LD, though there are stochastic fluctuations in individual cases. There are two primary routes for selection to increase LD [8]. The first is a hitchhiking effect (also known as a selective sweep), in which an entire haplotype with an advantageous variant is rapidly selected to high frequency or even fixation [8,15], leading to a high degree of LD carried by the selected haplotype. This occurs in the process of positive (adaptive) selection. Purifying (negative) selection against deleterious variants can also increase LD, as the deleterious haplotypes are swept from

Table 1: Recombination rates for highly and less highly conserved regions

	Recombination rate (cM/Mb)
Highly conserved regions	2.05
Less highly conserved regions	2.24
Highly conserved non-coding regions	2.54
Less highly conserved non-coding regions	2.98
Highly conserved coding rich-regions	0.62
Less highly conserved coding rich-regions	1.46
Small genes	1.78

These regions correspond to the regions in Figure 1D, in which we used highly and less highly conserved regions, highly and less highly non-coding conserved regions, and regions enriched (>20% in the bases) with highly and less highly conserved coding regions. As in Figure 1D, we selected only regions where the proportion of repeats was <20%, and since after this adjustment we found outliers of LD related to extreme GC-content, we further selected regions where the GC-content was 45–65%. For reference, we list the recombination rate in small genes with sizes up to 1000 bps and with the same conditions as to repeat proportion and GC-content. The average recombination rate in the genome was 1.33 cM/Mb.

the population [8]. The second route is epistatic selection for combinations of alleles at multiple loci [6,8], in which natural selection may favor or may not favor certain combinations of alleles that work synergistically. Recent studies [1,5,7] on LD patterns by large-scale genotyping datasets have demonstrated that LD in genic regions is strong at sizes roughly up to 100 to 200 kb. Because most genes are exposed to purifying (not positive) selection, these studies illustrate that the overall effect of purifying selection is to increase the degree of LD. This in turn suggests that the degree of LD may be strong in regions that are evolutionarily conserved between distantly related species, e.g., humans and mice, because it is evolutionarily conserved regions that remain unchanged by purifying selection [16].

However, a recent HapMap companion paper [9] reported that, although base-pairs in regions conserved between the human and mouse genomes were associated with low LD when sequence features were analyzed individually, the sequence conservation was not identified as an important predictor of LD in a multiple linear regression analysis. Consistent with these findings, although we initially found that conserved regions showed low LD levels, after consideration of sequence features one-by-one, we eventually determined that this relationship between LD and conserved regions was distorted by the lack of repetitive elements in such regions.

Thus, we excluded the effect of repetitive elements (and GC-content as well) in order to determine the relationship between LD and conserved regions. However, the result was not as simple as expected. We did not see a strong association between the level of LD and the degree of conservation in overall conserved regions but observed that strong LD was related to strong conservation in conserved coding regions. In addition, the LD level was not strongly related to sequence conservation in conserved non-coding regions. Because it has been demonstrated that even in non-coding regions, conserved regions include more functionally important segments, such as regulatory elements, than non-conserved regions [17,18], conservation thus seems to indicate selective constraint even in non-coding regions. Taking this into account, one interpretation of our results is that selective force works differentially between coding and non-coding regions. Purifying selection works on the function of exons' final protein products and may not allow frequent recombination between sequential series of coding sequences, which leads to strong LD in these sequences. Meanwhile, the similar LD levels in highly and less highly conserved non-coding regions may be explained by the independence of functional units, such as regulatory elements, present in those non-coding regions. That is, just individual alleles in conserved non-coding regions may be exposed to selective pressure; therefore, they may more often accept recombination between them. Alternatively, our results may suggest that conservation does not indicate selective constraint only for non-coding regions since non-coding regions might include too much noise unlike coding-regions, which are by definition functional.

Conclusion

Following the previous hypothesis that purifying selection increases the extent of LD, we examined whether LD was actually lower in evolutionarily conserved regions and attempted, by considering one potential factor at a time, to determine if a third factor may distort LD in conserved regions. We found that this tendency was associated with a lack of repetitive elements in those regions. We then showed that after correcting for the effect of repeat abundance, the degree of conservation itself was not strongly associated with the extent of LD in non-coding regions, but it was associated with LD in coding regions, which suggested that the effect of purifying selection on LD was more complex than expected from the previous hypothesis. This can be explained by the idea that natural selection works on the function of conserved exons' final protein products, while it works independently on the constituent alleles of conserved functional units in non-coding regions. In summary, purifying selection may prominently increase the extent of LD only when regions between alleles contain sequentially meaningful segments, such as segments translated into proteins. As we demonstrated, in-depth analyses are needed to elucidate the relationship between LD and sequence features. By means of such analyses, the LD patterns of the human genome may help to clarify the biological processes of recombination, mutation, and natural selection during the evolutionary history of human populations.

Methods

Detection of conserved regions

We downloaded the human (build 34) and mouse (build 32) genomic sequences from NCBI and followed a previously described procedure [19] to identify orthologous regions of the human and mouse genomes. We used BlastZ [19] with parameters C = 2, T = 1, K = 3000 to align human and mouse genomic sequences, and obtained alignments between the human and mouse. For overlapping alignment regions, we formed single contiguous regions. For example, when three regions started from coordinates 1 to 10, 5 to 15, and 20 to 30, respectively, we merged the former two regions into one region that started from 1 to 15. We used these contiguous alignment regions as conserved regions, which occupy 44% (1,255,655,305/2,818,767,476 bases, excluding the Y chromosome and gaps between contigs) of the human genome, which is almost the same percent (roughly 40%) as that detected by the previous study [20] of the mouse genome using BlastZ. For other genomic regions described below, we also formed single contiguous regions from any overlapping regions.

Plot of LD versus distance

To calculate LD, we downloaded genotype data (Release 16a) from the International HapMap Project [4] website. The samples were derived from 90 individuals in Utah, USA, from the Centre d'Etude du Polymorphisme Humain collection (CEU); from 45 Han Chinese in Beijing, China (CHB); from 44 Japanese in Tokyo, Japan (JPT); and from 90 Yoruba in Ibadan, Nigeria (YRI). The datasets of the four groups were treated independently to calculate linkage disequilibrium (LD). For each population, we considered a bi-allelic SNP to be validated in LD calculation as follows: 1) if the genotype data showed no significant deviation from Hardy-Weinberg equilibrium (Fisher's exact $p > 0.001$); 2) if the minor-allele frequency was greater than 0.2; 3) if the Mendelian inconsistency equaled zero; 4) if the position was not found at multiple chromosomal locations; 5) if the position was not located within a repeat element.

Using Haploview [21], we calculated pairwise r^2 and $|D'|$ for all possible pairs of validated SNPs that were separated by distances of less than 100 kb on the same contigs. We selected r^2 (and $|D'|$) values of SNP pairs both located within specified regions (e.g., conserved regions), and placed them into window bins according to predetermined ranges of distance between SNPs [6]. For a data point at position x in the plot figure (Fig. 1), we set the range of each sliding window from $k^{(-1/2)}x$ to $k^{(1/2)}x$ (corresponding to the range from $\log_{10}x-(1/2)\log_{10}k$ to $\log_{10}x+(1/2)\log_{10}k$ on a log scale), where $k = 1.5$, and we set the data point of the next sliding window at position lx (corresponding to $\log_{10}x+\log_{10}l$ on a log scale), where $l = 1.1$. Within each sliding window, we calculated the frequency of complete or nearly complete LD ($r^2 > 0.8$, $|D'| > 0.9$), and plotted the data point only when the sample size of those LD values was 100 or more (Fig. 1A, B and C), or 10 or more (Fig. 1D). In permutation tests, we randomly shuffled LD values of conserved and non-conserved regions (for each of the 0–10, 10–20, 20–30, 30–40 kb windows) 10,000 times, and each time we calculated a ratio of the two strong LD fractions of randomized conserved and non-conserved regions to get a p-value.

Datasets intersected with large-scale genomic features

We extracted gene positions from the NCBI Build 34 seq_gene.md mapview annotation file and created datasets consisting of conserved/non-conserved regions intersected with genic/non-genic regions. We also produced datasets involving conserved/non-conserved regions intersected with telomeric, centromeric, and other residual genomic regions. Because it is difficult to strictly define telomeric or centromeric regions, for this intersection we used two definitions (5% or 10%) of distances from the ends of the chromosomal arms distal and proximal to the centromere.

Highly conserved regions

We downloaded a table (mouse net table) containing coordinates of conserved regions between the human and mouse genomes from the UCSC genome browser website to define highly and less highly conserved regions. We iteratively adjusted a score parameter that was proportional to the sequence identity to obtain two sets of conserved regions that occupied approximately 5% and 40% of the genome, corresponding respectively to highly conserved regions and less highly conserved regions [20].

Regression analysis and permutation tests

We undertook a regression analysis to evaluate quantitatively how LD is influenced by a given sequence feature for each population. We regressed r^2 values adjusted by the physical distances with base-pair proportions of each sequence feature within SNP pairs, irrespective of categories of conserved and non-conserved regions. We first regressed observed r^2 values with the observed physical distances between SNPs using a model explicitly dependent on the distance as described below and obtained r^2 values that were expected from the distances:

$$E(r^2) = \sum_{n=0}^{3} a_n \cdot l^n,$$

where l and a_n are the physical distance and the regression coefficients (we used only SNP pairs with distances of 10 k to 100 k bps). We used Akaike's Information Criteria (AIC) to determine that this model was the best fit among several simple (linear, exponential, logarithmic, power, quadratic, or cubic) models.

To adjust the effect of physical distance on LD, we calculated the residual (r^2_{res}) by subtracting the expected r^2 value from the observed r^2 value,

$$r^2_{res} = r^2 - E(r^2),$$

which we regressed with the observed feature proportion,

$$r^2_{res} = cp + d,$$

where p, c, and d are the observed proportion, the regression coefficient, and the intercept, respectively. This coefficient was used to compare the influence of each feature on LD (see Additional file 2). We applied this simple regression to each sequence feature instead of multivariate regression, since simple regression is widely considered more effective for interpreting the regression coefficient.

In the simple regression, we obtained a further residual for a permutation test. In this test, we randomly shuffled the residuals of conserved and non-conserved categories and obtained a difference between the means of the residuals over the two artificial categories 10,000 times to calculate a p-value.

Partial correlation analysis

We performed a partial correlation analysis to simultaneously evaluate the effects of multiple sequence features on LD, which cannot be attained simply by plotting LD. We used two partial correlation coefficients (R_1 and R_2) between r^2 and the base-pair proportion (p_{cns}) of conserved regions within SNP pairs, given only physical distance (l); and given both physical distance and the proportion ($p_{feature}$) of each of the sequence features, such as GC, gene, or repeat:

$$R_1(r^2, p_{cns} \mid l),$$

$$R_2(r^2, p_{cns} \mid l, p_{feature}).$$

If the value of R_1 differed from that of R_2, we attributed the difference to $p_{feature}$.

Recombination rates

We downloaded the datasets of fine-scale recombination rates from the HapMap Project [4] website. We calculated the average of recombination rates across specified regions: $(\Sigma\rho_i)/l$, where l is the total length of the regions and ρ_i is a recombination rate at a base position i within the regions.

Authors' contributions

MK and TaT planned the research. MK performed the analyses and wrote the manuscript. TAJ, ToT, AS, YO, YN, and TaT reviewed the manuscript. All authors read and approved the final manuscript.

Additional material

Additional File 1

A figure showing plots of the moving average of the fraction of complete or nearly complete LD ($r^2 > 0.8$) versus distance between SNPs for all four populations.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-326-S1.pdf]

Additional File 2

A table showing the regression coefficients, ratios of base pairs in conserved regions to non-conserved regions, and base-pair fractions in the genome for sequence features.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-326-S2.doc]

Additional File 3

A table showing the results of partial correlation analysis to detect sequence features that involve weaker LD in conserved regions.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-326-S3.doc]

Acknowledgements

We thank T. Kawaguchi, H. Kawakami, E. Okutsu, and T. Morizono for writing some of the computer programs used in the analyses. TaT was supported by Takeda Science Foundation.

References

1. Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, Dibling T, Tinsley E, Kirby S, Carter D, Papaspyridonos M, Livingstone S, Ganske R, Lohmussaar E, Zernant J, Tonisson N, Remm M, Magi R, Puurand T, Vilo J, Kurg A, Rice K, Deloukas P, Mott R, Metspalu A, Bentley DR, Cardon LR, Dunham I: **A first-generation linkage disequilibrium map of human chromosome 22.** *Nature* 2002, **418(6897):**544-548.
2. Tomlinson IP, Bodmer WF: **The HLA system and the analysis of multifactorial genetic disease.** *Trends Genet* 1995, **11(12):**493-498.
3. Chakravarti A, Phillips JA 3rd, Mellits KH, Buetow KH, Seeburg PH: **Patterns of polymorphism and linkage disequilibrium suggest independent origins of the human growth hormone gene cluster.** *Proc Natl Acad Sci U S A* 1984, **81(19):**6085-6089.
4. The International HapMap Consortium: **The International HapMap Project.** *Nature* 2003, **426(6968):**789-796.
5. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR: **Whole-genome patterns of common DNA variation in three human populations.** *Science* 2005, **307(5712):**1072-1079.
6. Tsunoda T, Lathrop GM, Sekine A, Yamada R, Takahashi A, Ohnishi Y, Tanaka T, Nakamura Y: **Variation of gene-based SNPs and linkage disequilibrium patterns in the human genome.** *Hum Mol Genet* 2004, **13(15):**1623-1632.
7. The International HapMap Consortium: **A haplotype map of the human genome.** *Nature* 2005, **437(7063):**1299-1320.
8. Ardlie KG, Kruglyak L, Seielstad M: **Patterns of linkage disequilibrium in the human genome.** *Nat Rev Genet* 2002, **3(4):**299-309.
9. Smith AV, Thomas DJ, Munro HM, Abecasis GR: **Sequence features in regions of weak and strong linkage disequilibrium.** *Genome Res* 2005, **15(11):**1519-1534.
10. Wall JD, Pritchard JK: **Haplotype blocks and linkage disequilibrium in the human genome.** *Nat Rev Genet* 2003, **4(8):**587-597.
11. Tishkoff SA, Verrelli BC: **Patterns of human genetic diversity: implications for human evolutionary history and disease.** *Annu Rev Genomics Hum Genet* 2003, 4:293-340.
12. Phillips MS, Lawrence R, Sachidanandam R, Morris AP, Balding DJ, Donaldson MA, Studebaker JF, Ankener WM, Alfisi SV, Kuo FS, Camisa AL, Pazorov V, Scott KE, Carey BJ, Faith J, Katari G, Bhatti HA, Cyr JM, Derohannessian V, Elosua C, Forman AM, Grecco NM, Hock CR, Kuebler JM, Lathrop JA, Mockler MA, Nachtman EP, Restine SL, Varde SA, Hozza MJ, Gelfand CA, Broxholme J, Abecasis GR, Boyce-Jacino MT, Cardon LR: **Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots.** *Nat Genet* 2003, **33(3):**382-387.
13. Chiaromonte F, Yang S, Elnitski L, Yap VB, Miller W, Hardison RC: **Association between divergence and interspersed repeats in mammalian noncoding genomic DNA.** *Proc Natl Acad Sci U S A* 2001, **98(25):**14503-14508.
14. Boissinot S, Entezam A, Young L, Munson PJ, Furano AV: **The insertional history of an active family of L1 retrotransposons in humans.** *Genome Res* 2004, **14(7):**1221-1231.
15. Abecasis GR, Ghosh D, Nichols TE: **Linkage disequilibrium: ancient history drives the new genetics.** *Hum Hered* 2005, **59(2):**118-124.

16. Ruvolo M: **Comparative primate genomics: the year of the chimpanzee.** *Curr Opin Genet Dev* 2004, **14(6)**:650-656.
17. Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE: **Human-mouse genome comparisons to locate regulatory sites.** *Nat Genet* 2000, **26(2)**:225-228.
18. Pennacchio LA: **Insights from human/mouse genome comparisons.** *Mamm Genome* 2003, **14(7)**:429-436.
19. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W: **Human-mouse alignments with BLASTZ.** *Genome Res* 2003, **13(1)**:103-107.
20. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420(6915)**:520-562.
21. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps.** *Bioinformatics* 2005, **21(2)**:263-265.

J Hum Genet (2006) 51:1087–1099
DOI 10.1007/s10038-006-0070-9

ORIGINAL ARTICLE

Identification of a novel non-coding RNA, *MIAT*, that confers risk of myocardial infarction

Nobuaki Ishii · Kouichi Ozaki · Hiroshi Sato · Hiroya Mizuno · Susumu Saito · Atsushi Takahashi · Yoshinari Miyamoto · Shiro Ikegawa · Naoyuki Kamatani · Masatsugu Hori · Satoshi Saito · Yusuke Nakamura · Toshihiro Tanaka

Received: 14 August 2006 / Accepted: 31 August 2006 / Published online: 26 October 2006

Abstract Through a large-scale case-control association study using 52,608 haplotype-based single nucleotide polymorphism (SNP) markers, we identified a susceptible locus for myocardial infarction (MI) on chromosome 22q12.1. Following linkage disequilibrium (LD) mapping, haplotype analyses revealed that six SNPs in this locus, all of which were in complete LD, showed markedly significant association with MI (χ^2= 25.27, P=0.0000005; comparison of allele frequency, 3,435 affected individuals versus 3,774 controls, in the case of intron 1 5,338 C>T; rs2331291). Within this locus, we isolated a complete cDNA of a novel gene, designated myocardial infarction associated transcript (*MIAT*). *MIAT* has five exons, and in vitro translation assay showed that *MIAT* did not encode any translational product, indicating that this is likely to be a functional RNA. In vitro functional analyses revealed that the minor variant of one SNP in exon 5 increased transcriptional level of the novel gene. Moreover, unidentified nuclear protein(s) bound more intensely to risk allele than non-risk allele. These results indicate that the altered expression of *MIAT* by the SNP may play some role in the pathogenesis of MI.

Keywords Case-control association study · Myocardial infaraction associated transcript · Novel gene · SNP analysis

Introduction

Coronary artery diseases (CAD) including myocardial infarction (MI) have been the major cause of mortality and morbidity among late-onset diseases in many industrialized countries with a Western lifestyle (Breslow 1997; Braunwald 1997). MI often occurs without any preceding clinical signs, and is followed by severe complications, especially ventricular fibrillation and cardiac rupture which result in sudden death. Although recent advances in the treatment and diagnosis of MI have improved the quality of life for MI patients, its morbidity is still high.

Epidemiological studies revealed a variety of coronary risk factors, including type 2 diabetes mellitus,

N. Ishii · K. Ozaki · T. Tanaka (✉)
Laboratory for Cardiovascular Diseases,
SNP Research Center, The Institute of Physical and
Chemical Research (RIKEN), 4-6-1 Shirokanedai,
Minato-ku, Tokyo 108-8639, Japan
e-mail: toshitan@ims.u-tokyo.ac.jp

N. Ishii · Satoshi Saito
Division of Cardiology, Department of Medicine,
Nihon University School of Medicine, Tokyo, Japan

H. Sato · H. Mizuno · M. Hori
Department of Cardiovascular Medicine,
Osaka University Graduate School of Medicine,
Suita, Japan

Susumu Saito · Y. Nakamura
Laboratory for Genotyping, SNP Research Center,
The Institute of Physical and Chemical Research (RIKEN),
Tokyo, Japan

A. Takahashi · N. Kamatani
Laboratory for Statistical Analysis, SNP Research Center,
The Institute of Physical and Chemical Research (RIKEN),
Tokyo, Japan

Y. Miyamoto · S. Ikegawa
Laboratory for Bone and Joint Disease,
SNP Research Center, The Institute of Physical and
Chemical Research (RIKEN), Tokyo, Japan

hypercholesterolemia, hypertension, and obesity. There are also studies reporting a genetic factor of this disorder; one reported first-degree relatives of patients who have had an acute MI prior to age 55 years have 2–7 times higher risk of MI (Lusis et al. 2004). A twin study indicated an eight-fold increase in risk of death from MI when a first twin dies of MI before age 55 years (Marenberg et al. 1994). In this context, common genetic variants are considered to contribute to genetic risks of common diseases (Lander 1996; Risch and Merikangas 1996; Collins et al. 1997).

To date, various genetic variants that confer susceptibility to MI have been indicated to be present on several chromosomal loci through linkage analyses or case-control association studies using single nucleotide polymorphism (SNP) (Topol et al. 2001; Yamada et al. 2002; Ozaki et al. 2002, 2004, 2006; Stenina et al. 2003; Helgadottir et al. 2004). Case-control association study by means of genome-wide SNP analysis is one of the most powerful approaches to identify genetic variants susceptibility to common diseases.

We report here identification of SNPs in myocardial infarction associated transcript (*MIAT*) that were associated with MI through our comprehensive SNP association study. We also demonstrated the possible transcriptional effect of this variation on the expression level of *MIAT*.

Materials and methods

DNA samples

This study included 3,464 Japanese individuals with myocardial infarction who were referred to Osaka Acute Coronary Insufficiency Study Group, which involved the cardiovascular units of 25 hospitals in Osaka. The diagnosis of definite myocardial infarction has been previously described (Ohnishi et al. 2000; Ozaki et al. 2002). The control individuals consisted of 3,819 members of the general population who were recruited through several medical institutes in Japan. DNAs were prepared from these samples according to standard protocols. All individuals were Japanese, gave written informed consent to participate in the study, or their parents gave them when they were under 20 years old, according to the process approved by the relevant Ethical Committee at SNP Research Center, The Institute of Physical and Chemical Research (RIKEN) Yokohama.

SNP discovery and genotyping

Protocols for PCR primers, PCR experiments, DNA extraction, DNA sequencing and genotyping of SNPs have been previously described (Iida et al. 2001). For SNP discovery, genomic DNAs from 24 Japanese individuals were used, and direct sequencing was performed using capillary sequencer (ABI3700; Applied Biosystems, Foster City, Calif., USA).

Statistical analysis

We assessed association and Hardy–Weinberg equilibrium by χ^2-test (Yamada et al. 2001; Ozaki et al. 2002). Haplotype block and haplotype frequencies were estimated using SNPAlyze software (DYNACOM, Chiba, Japan) and Haploview v3.2 (Barrett et al. 2005).

Northern blot analysis

Human multiple-tissue Northern (MTN) blots (Clontech, Palo Alto, Calif., USA) were pre-hybridized and hybridized with α-[^{32}P]-dCTP labeled cDNA fragment prepared by PCR using primer pair shown in Table 1, as a probe. Washed membranes were autoradiographed for 7 days at –80°C.

Isolation of full-length cDNA

A human fetal brain cDNA library was constructed with combination of gene specific linker primers, random hexamer linker primer and oligo(dT) linker primer using ZAP cDNA synthesis kit (Stratagene, La Jolla, Calif., USA) according to the manufacturer's protocol. The library was screened with the same probe as Northern experiment. Positive clones were selected and their insert cDNAs were excised in vivo in pBluescriptIISK(-) (Stratagene) according to the manufacturer's protocol. To obtain the missing 5′- or 3′-portion, we performed a rapid amplification of cDNA ends (RACE) using BD SMART RACE cDNA Amplification Kit (Clontech) according to the manufacturer's instructions. Primers for full-length cDNA isolation were shown in Table 1.

In vitro translation assay

Four kinds of plasmids corresponding to variant 1–4, which were obtained through screening of ZAP human fetal brain cDNA library for isolation of *MIAT*