ープとして参考にして祖先型配列を導く必要があろう。本研究で見いだされた転写開始点の種差と、5'UTR での特徴的な indel の入り方は、ヒトとチンパンジーの遺伝子発現制御の違いに関連するとも考えられる。したがって、indel を含めた全長配列の比較解析は、遺伝子機能解析上も重要な情報をもたらすことを確認した。なお、本年度は、ヒトにおいて心血管系異常ならびに肥満に関連すると考えられている遺伝子のホモログ（心血管疾患関連３９遺伝子[１２２クローン]、肥満関連３１遺伝子[１０６クローン]の塩基配列決定もすすめた。

E．結論

ヒトの遺伝子機能の解析に、比較ゲノム学的視点が次第に重要となってきた。ヒトと最も近縁のチンパンジーについて、全長 cDNA を対象にした研究が重要な情報をもたらすものと考えられる。本研究では、チンパンジー遺伝子の全長配列を解読し、ヒトの対応する配列と比較することにより、転写開始点・終了点の特定、塩基置換ならびに挿入・欠失の比較と発現への影響などの遺伝子の機能に結びつく情報を得た。これらの情報は、今後のマイクロアレイや定量的 PCR 法、さらにはプロテオミクスによる発現解析を行ううえで、有用と考える。また、ヒトの翻訳領域における SNP の意義などを調べるうえでも、チンパンジーの配列情報が価値の高い研究資源となると考えられるので、その供給のための作業をさらに進めたい。

F．研究発表

１．論文発表

1) K. Nagao, N. Takenaka, M, Hirai and S. Kawamura: Coupling and decoupling of evolutionary mode between X- and Y-chromosomal red-green opsin genes in owl monkeys. Gene (2005) in press.

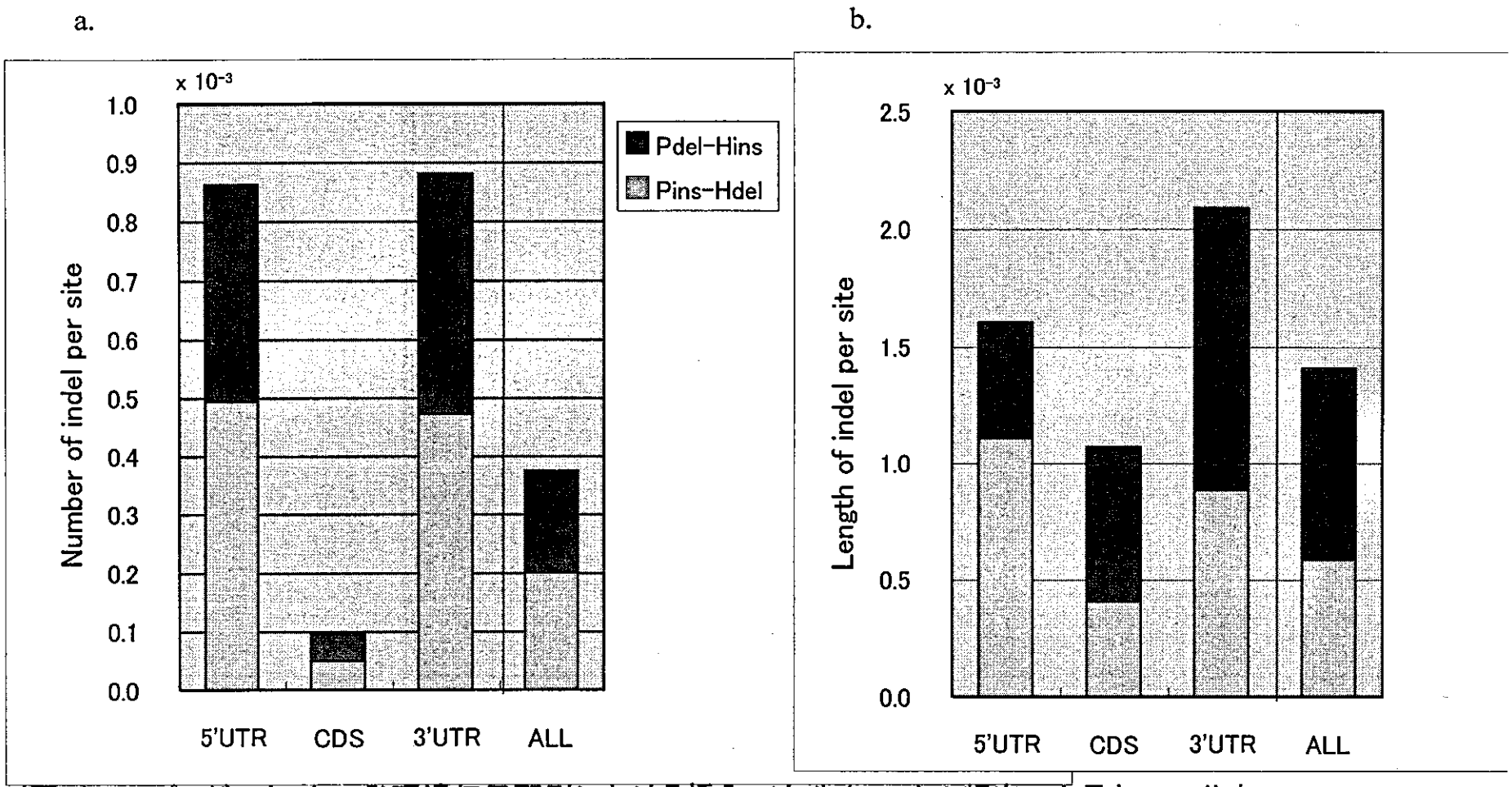
２．学会発表

1) Y. Suto , M.H. Park, D.H. Seo, M. Hammond, E. Smart, R. Matsuoka, M. Hirai, T. Juji and Y. Ishikawa : Fiber fluorescence in situ hybridization (FISH) analysis of Rhesus (Rh) blood group locus. Human Genome Variation Society Scientific &Annual General Meeting 2004 (26 October, Toronto, Canada)

2）坂手龍一、今西規、平井百樹、橋本雄之、五条堀孝：ヒトとチンパンジーのcDNA比較解析から探るヒト遺伝子の進化。日本遺伝学会第７６回大会。９．２７- ２９（大阪大学）。

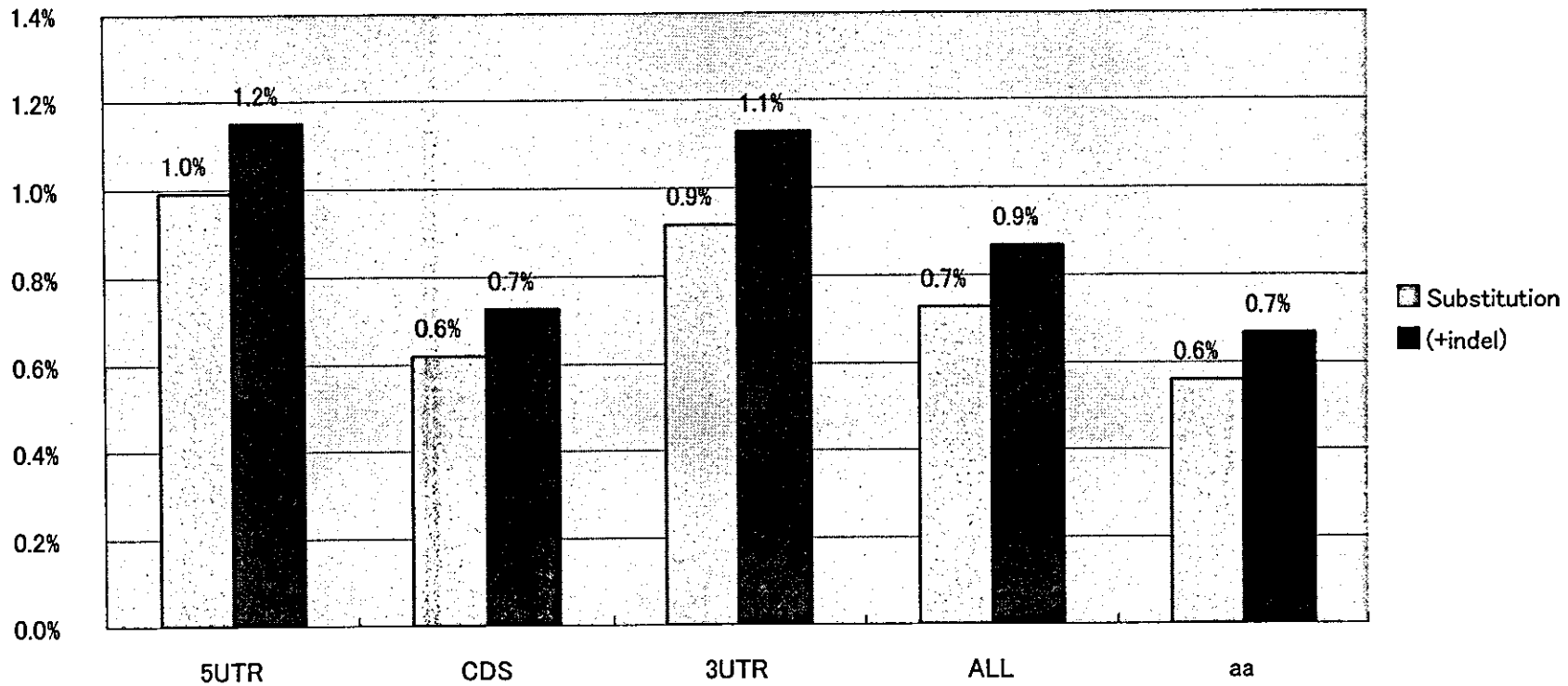
G．知的所有権の取得状況

取得なし



(図1)チンパンジーとヒトの発現遺伝子配列における挿入／欠失(indel)の頻度(a)と長さ(b)の分布(per nucleotide site)。どちらの配列での挿入(欠失)かの内訳も示す。
Pdel-Hins: チンパンジー配列で欠失(ヒト配列で挿入)　Pins-Hdel: チンパンジー配列で挿入(ヒト配列で欠失)

Indelの長さに注目すると(図1−b)、5'UTRではPins-Hdel がPdel-Hinsより比率が高く、CDSでは逆にPdel-Hins が Pins-Hdelより比率が高い(Fisher's exact test, $P < 0.05$)。



（図2） チンパンジー全長cDNA配列とヒトの対応配列との領域別比較解析
（左側）塩基置換率　（右側）挿入／欠失を算入した場合の種差

表 1　全長解析を行ったチンパンジーのクローン

clone	RefSeq_acc	RefSeq_definition
PccB0268	NM_052820	Homo sapiens coronin, actin binding protein, 2A (CORO2A), transcript variant 2, mRNA.
PccB0407	NM_001814	Homo sapiens cathepsin C (CTSC), transcript variant 1, mRNA.
PccB0430	NM_001967	Homo sapiens eukaryotic translation initiation factor 4A, isoform 2 (EIF4A2), mRNA.
PccB0478	NM_201541	Homo sapiens NDRG family member 2 (NDRG2), transcript variant 8, mRNA.
PccB0660	NM_006623	Homo sapiens phosphoglycerate dehydrogenase (PHGDH), mRNA.
PccB0720	NM_002136	Homo sapiens heterogeneous nuclear ribonucleoprotein A1 (HNRPA1), transcript variant 1, mRNA.
PccB0722	NM_139207	Homo sapiens nucleosome assembly protein 1-like 1 (NAP1L1), transcript variant 1, mRNA.
PccB0730	NM_005165	Homo sapiens aldolase C, fructose-bisphosphate (ALDOC), mRNA.
PccB0759	NM_004046	Homo sapiens ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit, isoform 1, cardiac muscle (ATP5A1), nuclear gene encoding mitochondrial protein, transcript variant 2, mRNA.
PccB0771	NM_014741	Homo sapiens KIAA0652 gene product (KIAA0652), mRNA.
PccB0922	NM_001101	Homo sapiens actin, beta (ACTB), mRNA.
PccB0929	NM_001918	Homo sapiens dihydrolipoamide branched chain transacylase E2 (DBT), mRNA.
PccB1124	NM_001402	Homo sapiens eukaryotic translation elongation factor 1 alpha 1 (EEF1A1), mRNA.
PccB1539	NM_000371	Homo sapiens transthyretin (prealbumin, amyloidosis type I) (TTR), mRNA.
PccB1539	NM_005165	Homo sapiens aldolase C, fructose-bisphosphate (ALDOC), mRNA.
PccB1550	NM_005184	Homo sapiens calmodulin 3 (phosphorylase kinase, delta) (CALM3), mRNA.
PccB1704	NM_021999	Homo sapiens integral membrane protein 2B (ITM2B), mRNA.
PccB1740	NM_000709	Homo sapiens branched chain keto acid dehydrogenase E1, alpha polypeptide (maple syrup urine disease) (BCKDHA), mRNA.
PccB1795	NM_000016	Homo sapiens acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain (ACADM), nuclear gene encoding mitochondrial
PccB1949	NM_000194	Homo sapiens hypoxanthine phosphoribosyltransferase 1 (Lesch-Nyhan syndrome) (HPRT1), mRNA.
PccB2065	NM_022132	Homo sapiens methylcrotonoyl-Coenzyme A carboxylase 2 (beta) (MCCC2), mRNA.
PccB2102	NM_022477	Homo sapiens NDRG family member 3 (NDRG3), transcript variant 2, mRNA.
PccB2111	NM_002435	Homo sapiens mannose phosphate isomerase (MPI), mRNA.
PccB2143	NM_006082	Homo sapiens tubulin, alpha, ubiquitous (K-ALPHA-1), mRNA.
PccB2156	NM_130811	Homo sapiens synaptosomal-associated protein, 25kDa (SNAP25), transcript variant 2, mRNA.
PccB2387	NM_021999	Homo sapiens integral membrane protein 2B (ITM2B), mRNA.
PccB2677	NM_000165	Homo sapiens gap junction protein, alpha 1, 43kDa (connexin 43) (GJA1), mRNA.
PccB2822	NM_181598	Homo sapiens spastic paraplegia 3A (autosomal dominant) (SPG3A), mRNA.
PccB2891	NM_022808	Homo sapiens small nuclear ribonucleoprotein polypeptide N (SNRPN), transcript variant 5, mRNA.
PccB3027	NM_002107	Homo sapiens H3 histone, family 3A (H3F3A), mRNA.
PccB3354	NM_201414	Homo sapiens amyloid beta (A4) precursor protein (protease nexin-II, Alzheimer disease) (APP), transcript variant 3, mRNA.
PccB3405	NM_000968	Homo sapiens ribosomal protein L4 (RPL4), mRNA.
PccB3630	NM_003754	Homo sapiens eukaryotic translation initiation factor 3, subunit 5 epsilon, 47kDa (EIF3S5), mRNA.

PccB3639	NM_003908	Homo sapiens eukaryotic translation initiation factor 2, subunit 2 beta, 38kDa (EIF2S2), mRNA.
PccB4236	NM_184041	Homo sapiens aldolase A, fructose-bisphosphate (ALDOA), transcript variant 2, mRNA.
PccB4238	NM_002136	Homo sapiens heterogeneous nuclear ribonucleoprotein A1 (HNRPA1), transcript variant 1, mRNA.
PccB4243	NM_201414	Homo sapiens amyloid beta (A4) precursor protein (protease nexin-II, Alzheimer disease) (APP), transcript variant 3, mRNA.
PccB4252	NM_001063	Homo sapiens transferrin (TF), mRNA.
PccB4550	NM_183079	Homo sapiens prion protein (p27-30) (Creutzfeld-Jakob disease, Gerstmann-Strausler-Scheinker syndrome, fatal familial insomnia) (PRNP), transcript variant 2, mRNA.
PflB0128	NM_153649	Homo sapiens tropomyosin 3 (TPM3), mRNA.
PflB0280	NM_001873	Homo sapiens carboxypeptidase E (CPE), mRNA.
PflB0380	NM_015292	Homo sapiens likely ortholog of mouse membrane bound C2 domain containing protein (MBC2), mRNA.
PflB0650	NM_001554	Homo sapiens cysteine-rich, angiogenic inducer, 61 (CYR61), mRNA.
PflB1716	NM_002047	Homo sapiens glycyl-tRNA synthetase (GARS), mRNA.
PflB2127	NM_005822	Homo sapiens Down syndrome critical region gene 1-like 1 (DSCR1L1), mRNA.
PflB2865	NM_201414	Homo sapiens amyloid beta (A4) precursor protein (protease nexin-II, Alzheimer disease) (APP), transcript variant 3, mRNA.
PflB2870	NM_000365	Homo sapiens triosephosphate isomerase 1 (TPI1), mRNA.
PflB2983	NM_001539	Homo sapiens DnaJ (Hsp40) homolog, subfamily A, member 1 (DNAJA1), mRNA.
PflB3550	NM_004331	Homo sapiens BCL2/adenovirus E1B 19kDa interacting protein 3-like (BNIP3L), mRNA.
PflB3724	NM_203417	Homo sapiens Down syndrome critical region gene 1 (DSCR1), transcript variant 2, mRNA.
PflB3863	NM_030940	Homo sapiens HESB like domain containing 2 (HBLD2), mRNA.
PflB4078	NM_000980	Homo sapiens ribosomal protein L18a (RPL18A), mRNA.
PflB5542	NM_022975	Homo sapiens fibroblast growth factor receptor 2 (bacteria-expressed kinase, keratinocyte growth factor receptor, craniofacial dysostosis 1, Crouzon syndrome, Pfeiffer syndrome, Jackson-Weiss syndrome) (FGFR2), transcript variant 8, mR十。
PflB7187	NM_000430	Homo sapiens platelet-activating factor acetylhydrolase, isoform Ib, alpha subunit 45kDa (PAFAH1B1), mRNA.
PflB7643	NM_201414	Homo sapiens amyloid beta (A4) precursor protein (protease nexin-II, Alzheimer disease) (APP), transcript variant 3, mRNA.
PflB7895	NM_001693	Homo sapiens ATPase, H+ transporting, lysosomal 56/58kDa, V1 subunit B, isoform 2 (ATP6V1B2), mRNA.
PflB8491	NM_002480	Homo sapiens protein phosphatase 1, regulatory (inhibitor) subunit 12A (PPP1R12A), mRNA.
PflB9157	NM_019597	Homo sapiens heterogeneous nuclear ribonucleoprotein H2 (H') (HNRPH2), mRNA.
PflB9311	NM_002823	Homo sapiens prothymosin, alpha (gene sequence 28) (PTMA), mRNA.
PflB9342	NM_152640	Homo sapiens DCP1 decapping enzyme homolog B (S. cerevisiae) (DCP1B), mRNA.
PflB9547	NM_022758	Homo sapiens chromosome 6 open reading frame 106 (C6orf106), transcript variant 2, mRNA.
PorA0409	NM_006009	Homo sapiens tubulin, alpha 3 (TUBA3), mRNA.
PorA0413	NM_004396	Homo sapiens DEAD (Asp-Glu-Ala-Asp) box polypeptide 5 (DDX5), mRNA.
PorA0418	NM_004684	Homo sapiens SPARC-like 1 (mast9, hevin) (SPARCL1), mRNA.
PorA0443	NM_014478	Homo sapiens calcitonin gene-related peptide-receptor component protein (RCP9), mRNA.
PorA0450	NM_007177	Homo sapiens TU3A protein (TU3A), mRNA.
PorA0463	NM_001549	Homo sapiens interferon-induced protein with tetratricopeptide repeats 3 (IFIT3), mRNA.
PorA0465	NM_012286	Homo sapiens mortality factor 4 like 2 (MORF4L2), mRNA.

PorA0649	NM_003020	Homo sapiens secretory granule, neuroendocrine protein 1 (7B2 protein) (SGNE1), mRNA.
PorA0657	NM_001010942	Homo sapiens RAP1B, member of RAS oncogene family (RAP1B), transcript variant 2, mRNA.
PorA0666	NM_001677	Homo sapiens ATPase, Na+/K+ transporting, beta 1 polypeptide (ATP1B1), transcript variant 1, mRNA.
PorA0680	NM_014394	Homo sapiens growth hormone inducible transmembrane protein (GHITM), mRNA.
PorA0688	NM_001549	Homo sapiens interferon-induced protein with tetratricopeptide repeats 3 (IFIT3), mRNA.
PorA0791	NM_000157	Homo sapiens glucosidase, beta; acid (includes glucosylceramidase) (GBA), transcript variant 1, mRNA.
PorA0806	NM_201428	Homo sapiens reticulon 3 (RTN3), transcript variant 2, mRNA.
PorA0834	NM_007008	Homo sapiens reticulon 4 (RTN4), transcript variant 3, mRNA.
PorA0842	NM_001469	Homo sapiens thyroid autoantigen 70kDa (Ku antigen) (G22P1), mRNA.
PorA0847	NM_005678	Homo sapiens SNRPN upstream reading frame (SNURF), transcript variant 1, mRNA.
PorA0848	NM_003295	Homo sapiens tumor protein, translationally-controlled 1 (TPT1), mRNA.
PorA0977	NM_002435	Homo sapiens mannose phosphate isomerase (MPI), mRNA.
PorA0990	NM_000990	Homo sapiens ribosomal protein L27a (RPL27A), mRNA.
PorA1107	NM_177924	Homo sapiens N-acylsphingosine amidohydrolase (acid ceramidase) 1 (ASAH1), transcript variant 1, mRNA.
PorA1118	NM_001908	Homo sapiens cathepsin B (CTSB), transcript variant 1, mRNA.
PorA1145	NM_000291	Homo sapiens phosphoglycerate kinase 1 (PGK1), mRNA.
PorA1157	NM_005678	Homo sapiens SNRPN upstream reading frame (SNURF), transcript variant 1, mRNA.
PorA1227	NM_019895	Homo sapiens chromosome 3 open reading frame 4 (C3orf4), mRNA.
PorA1323	NM_006888	Homo sapiens calmodulin 1 (phosphorylase kinase, delta) (CALM1), mRNA.
PorA1351	NM_006597	Homo sapiens heat shock 70kDa protein 8 (HSPA8), transcript variant 1, mRNA.
PorA1353	NM_201538	Homo sapiens NDRG family member 2 (NDRG2), transcript variant 5, mRNA.
PorA1369	NM_014306	Homo sapiens hypothetical protein HSPC117 (HSPC117), mRNA.
PorA1376	NM_000122	Homo sapiens excision repair cross-complementing rodent repair deficiency, complementation group 3 (xeroderma pigmentosum group B complementing) (ERCC3), mRNA.
PorA1393	NM_002079	Homo sapiens glutamic-oxaloacetic transaminase 1, soluble (aspartate aminotransferase 1) (GOT1), mRNA.
PorB0293	NM_006082	Homo sapiens tubulin, alpha, ubiquitous (K-ALPHA-1), mRNA.
PorB0308	NM_005348	Homo sapiens heat shock 90kDa protein 1, alpha (HSPCA), mRNA.
PorB0603	NM_000284	Homo sapiens pyruvate dehydrogenase (lipoamide) alpha 1 (PDHA1), mRNA.
PorB1179	NM_000146	Homo sapiens ferritin, light polypeptide (FTL), mRNA.
PorB1179	NM_000270	Homo sapiens nucleoside phosphorylase (NP), mRNA.
PorB1309	NM_001743	Homo sapiens calmodulin 2 (phosphorylase kinase, delta) (CALM2), mRNA.
PorB1332	NM_005348	Homo sapiens heat shock 90kDa protein 1, alpha (HSPCA), mRNA.
PorB1356	NM_001493	Homo sapiens GDP dissociation inhibitor 1 (GDI1), mRNA.
PorB1369	NM_004355	Homo sapiens CD74 antigen (invariant polypeptide of major histocompatibility complex, class II antigen-associated) (CD74), mRNA.
PorB1388	NM_203433	Homo sapiens Down syndrome critical region gene 2 (DSCR2), transcript variant 2, mRNA.
PorB1391	NM_012245	Homo sapiens SKI interacting protein (SKIIP), mRNA.
PorB1442	NM_005566	Homo sapiens lactate dehydrogenase A (LDHA), mRNA.

PorB1491	NM_001037	Homo sapiens sodium channel, voltage-gated, type I, beta (SCN1B), transcript variant a, mRNA.
PorB1614	NM_015665	Homo sapiens achalasia, adrenocortical insufficiency, alacrimia (Allgrove, triple-A) (AAAS), mRNA.
PstA0001	NM_133473	Homo sapiens zinc finger protein 431 (ZNF431), mRNA.
PstA0482	NM_001642	Homo sapiens amyloid beta (A4) precursor-like protein 2 (APLP2), mRNA.
PstA1183	NM_031314	Homo sapiens heterogeneous nuclear ribonucleoprotein C (C1/C2) (HNRPC), transcript variant 1, mRNA.
PstA1221	NM_002510	Homo sapiens glycoprotein (transmembrane) nmb (GPNMB), transcript variant 2, mRNA.
PstA1223	NM_003982	Homo sapiens solute carrier family 7 (cationic amino acid transporter, y+ system), member 7 (SLC7A7), mRNA.
PstA1554	NM_001428	Homo sapiens enolase 1, (alpha) (ENO1), mRNA.
PstA1604	NM_001630	Homo sapiens annexin A8 (ANXA8), mRNA.
PstA1671	NM_018479	Homo sapiens enoyl Coenzyme A hydratase domain containing 1 (ECHDC1), transcript variant 2, mRNA.
PstA1708	NM_002345	Homo sapiens lumican (LUM), mRNA.
PstA1724	NM_004862	Homo sapiens lipopolysaccharide-induced TNF factor (LITAF), mRNA.
PstA1929	NM_000423	Homo sapiens keratin 2A (epidermal ichthyosis bullosa of Siemens) (KRT2A), mRNA.
PstA1934	NM_005566	Homo sapiens lactate dehydrogenase A (LDHA), mRNA.
PstA1941	NM_000287	Homo sapiens peroxisomal biogenesis factor 6 (PEX6), mRNA.
PstA1943	NM_001101	Homo sapiens actin, beta (ACTB), mRNA.
PstA1947	NM_003380	Homo sapiens vimentin (VIM), mRNA.
PstA2015	NM_001264	Homo sapiens corneodesmosin (CDSN), mRNA.
PstA2019	NM_207514	Homo sapiens hypothetical protein FLJ20186 (FLJ20186), transcript variant 1, mRNA.
PstA2028	NM_000422	Homo sapiens keratin 17 (KRT17), mRNA.
PstA2236	NM_000709	Homo sapiens branched chain keto acid dehydrogenase E1, alpha polypeptide (maple syrup urine disease) (BCKDHA), mRNA.
PstA2336	NM_170707	Homo sapiens lamin A/C (LMNA), transcript variant 1, mRNA.
PstA2398	NM_001073	Homo sapiens UDP glycosyltransferase 2 family, polypeptide B11 (UGT2B11), mRNA.
PstA2422	NM_000700	Homo sapiens annexin A1 (ANXA1), mRNA.
PstA2450	NM_002139	Homo sapiens RNA binding motif protein, X-linked (RBMX), mRNA.
PstA2466	NM_016091	Homo sapiens eukaryotic translation initiation factor 3, subunit 6 interacting protein (EIF3S6IP), mRNA.
PstA2488	NM_001402	Homo sapiens eukaryotic translation elongation factor 1 alpha 1 (EEF1A1), mRNA.
PstA2497	NM_000183	Homo sapiens hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzyme A thiolase/enoyl-Coenzyme A hydratase (trifunctional protein), beta subunit (HADHB), mRNA.
PstA2528	NM_001010	Homo sapiens ribosomal protein S6 (RPS6), mRNA.
PstA2529	NM_153338	Homo sapiens hypothetical protein FLJ90165 (FLJ90165), mRNA.
PstA2561	NM_001064	Homo sapiens transketolase (Wernicke-Korsakoff syndrome) (TKT), mRNA.
PstA2589	NM_003016	Homo sapiens splicing factor, arginine/serine-rich 2 (SFRS2), mRNA.
PstA2598	NM_002277	Homo sapiens keratin, hair, acidic, 1 (KRTHA1), mRNA.
PstA2615	NM_001920	Homo sapiens decorin (DCN), transcript variant A1, mRNA.
PstA2616	NM_000424	Homo sapiens keratin 5 (epidermolysis bullosa simplex, Dowling-Meara/Kobner/Weber-Cockayne types) (KRT5), mRNA.
PstA2630	NM_000668	Homo sapiens alcohol dehydrogenase IB (class I), beta polypeptide (ADH1B), mRNA.

PstA2652	NM_006121	Homo sapiens keratin 1 (epidermolytic hyperkeratosis) (KRT1), mRNA.
PstA2662	NM_001424	Homo sapiens epithelial membrane protein 2 (EMP2), mRNA.
PstA2687	NM_000021	Homo sapiens presenilin 1 (Alzheimer disease 3) (PSEN1), transcript variant I-467, mRNA.
PstA2784	NM_006082	Homo sapiens tubulin, alpha, ubiquitous (K-ALPHA-1), mRNA.
PstA2806	NM_002277	Homo sapiens keratin, hair, acidic, 1 (KRTHA1), mRNA.
PstA3333	NM_005443	Homo sapiens 3'-phosphoadenosine 5'-phosphosulfate synthase 1 (PAPSS1), mRNA.
PstA3340	NM_007085	Homo sapiens follistatin-like 1 (FSTL1), mRNA.
PstA3420	NM_018418	Homo sapiens spermatogenesis associated 7 (SPATA7), mRNA.
PstA5207	NM_000424	Homo sapiens keratin 5 (epidermolysis bullosa simplex, Dowling-Meara/Kobner/Weber-Cockayne types) (KRT5), mRNA.
PstA5346	NM_002284	Homo sapiens keratin, hair, basic, 6 (monilethrix) (KRTHB6), mRNA.
PstA5383	NM_002278	Homo sapiens keratin, hair, acidic, 2 (KRTHA2), mRNA.
PstA5441	NM_005556	Homo sapiens keratin 7 (KRT7), mRNA.
PstA5446	NM_000261	Homo sapiens myocilin, trabecular meshwork inducible glucocorticoid response (MYOC), mRNA.
PstA5459	NM_001733	Homo sapiens complement component 1, r subcomponent (C1R), mRNA.
PstA5567	NM_181534	Homo sapiens keratin 25A (KRT25A), mRNA.
PstA5587	NM_000668	Homo sapiens alcohol dehydrogenase IB (class I), beta polypeptide (ADH1B), mRNA.
PstA5646	NM_001614	Homo sapiens actin, gamma 1 (ACTG1), mRNA.
PstA5667	NM_000158	Homo sapiens glucan (1,4-alpha-), branching enzyme 1 (glycogen branching enzyme, Andersen disease, glycogen storage disease type IV) (GBE1), mRN
PstA5737	NM_031412	Homo sapiens GABA(A) receptor-associated protein like 1 (GABARAPL1), mRNA.
PstA6018	NM_181534	Homo sapiens keratin 25A (KRT25A), mRNA.
PstA6246	NM_014741	Homo sapiens KIAA0652 gene product (KIAA0652), mRNA.
PstA6268	NM_001416	Homo sapiens eukaryotic translation initiation factor 4A, isoform 1 (EIF4A1), mRNA.
PstA6273	NM_000685	Homo sapiens angiotensin II receptor, type 1 (AGTR1), transcript variant 1, mRNA.
PstA6334	NM_002283	Homo sapiens keratin, hair, basic, 5 (KRTHB5), mRNA.
PstA6340	NM_004797	Homo sapiens adiponectin, C1Q and collagen domain containing (ADIPOQ), mRNA.
PstA6460	NM_199512	Homo sapiens steroid sensitive gene 1 (URB), transcript variant 2, mRNA.
PstA6552	NM_001967	Homo sapiens eukaryotic translation initiation factor 4A, isoform 2 (EIF4A2), mRNA.
PstA6602	NM_002300	Homo sapiens lactate dehydrogenase B (LDHB), mRNA.
PstA6629	NM_001788	Homo sapiens septin 7 (SEPT7), transcript variant 1, mRNA.
PstA6653	NM_003380	Homo sapiens vimentin (VIM), mRNA.
PstA6656	NM_000237	Homo sapiens lipoprotein lipase (LPL), mRNA.
PstA6761	NM_000421	Homo sapiens keratin 10 (epidermolytic hyperkeratosis; keratosis palmaris et plantaris) (KRT10), mRNA.
PstA6771	NM_175068	Homo sapiens keratin 6 irs3 (K6IRS3), mRNA.
PstA6831	NM_001154	Homo sapiens annexin A5 (ANXA5), mRNA.
PstA6843	NM_004168	Homo sapiens succinate dehydrogenase complex, subunit A, flavoprotein (Fp) (SDHA), nuclear gene encoding mitochondrial protein, mRNA.
PstA6902	NM_000095	Homo sapiens cartilage oligomeric matrix protein (COMP), mRNA.
PstA6914	NM_170708	Homo sapiens lamin A/C (LMNA), transcript variant 3, mRNA.

PstA7054	NM_021013	Homo sapiens keratin, hair, acidic, 4 (KRTHA4), mRNA.
PstA7147	NM_152843	Homo sapiens Hermansky-Pudlak syndrome 4 (HPS4), transcript variant 4, mRNA.
PstA7336	NM_000095	Homo sapiens cartilage oligomeric matrix protein (COMP), mRNA.
PstA7343	NM_000181	Homo sapiens glucuronidase, beta (GUSB), mRNA.
PstA7372	NM_006623	Homo sapiens phosphoglycerate dehydrogenase (PHGDH), mRNA.
PstA7438	NM_177924	Homo sapiens N-acylsphingosine amidohydrolase (acid ceramidase) 1 (ASAH1), transcript variant 1, mRNA.
PstA7445	NM_033059	Homo sapiens keratin associated protein 4-14 (KRTAP4-14), mRNA.
PstA7501	NM_004138	Homo sapiens keratin, hair, acidic, 3A (KRTHA3A), mRNA.
PstA7581	NM_002275	Homo sapiens keratin 15 (KRT15), mRNA.
PstA7590	NM_005557	Homo sapiens keratin 16 (focal non-epidermolytic palmoplantar keratoderma) (KRT16), mRNA.
PstA7623	NM_004071	Homo sapiens CDC-like kinase 1 (CLK1), mRNA.
PstA7631	NM_006597	Homo sapiens heat shock 70kDa protein 8 (HSPA8), transcript variant 1, mRNA.
PstA7636	NM_031961	Homo sapiens keratin associated protein 9-2 (KRTAP9-2), mRNA.
PstA7640	NM_000423	Homo sapiens keratin 2A (epidermal ichthyosis bullosa of Siemens) (KRT2A), mRNA.
PstA7657	NM_033448	Homo sapiens keratin 6 irs (KRT6IRS), mRNA.
PstA7678	NM_017566	Homo sapiens kelch domain containing 4 (KLHDC4), mRNA.
PstA7713	NM_001920	Homo sapiens decorin (DCN), transcript variant A1, mRNA.
PstA7915	NM_201414	Homo sapiens amyloid beta (A4) precursor protein (protease nexin-II, Alzheimer disease) (APP), transcript variant 3, mRNA.
PstA7974	NM_005572	Homo sapiens lamin A/C (LMNA), transcript variant 2, mRNA.
PstA8144	NM_006771	Homo sapiens keratin, hair, acidic, 8 (KRTHA8), mRNA.
PstA8148	NM_000165	Homo sapiens gap junction protein, alpha 1, 43kDa (connexin 43) (GJA1), mRNA.
PstA8159	NM_004484	Homo sapiens glypican 3 (GPC3), mRNA.
PtsC-10255	NM_002827	Homo sapiens protein tyrosine phosphatase, non-receptor type 1 (PTPN1), mRNA.
PtsC-11027	NM_033028	Homo sapiens Bardet-Biedl syndrome 4 (BBS4), mRNA.
PtsC-11175	NM_000147	Homo sapiens fucosidase, alpha-L- 1, tissue (FUCA1), mRNA.
PtsC-11774	NM_014567	Homo sapiens breast cancer anti-estrogen resistance 1 (BCAR1), mRNA.

平成１６年度　厚生労働科学研究費補助金（ヒトゲノム・再生医療等研究事業）

分　担　研　究　報　告　書

カニクイザルcDNAライブラリー作製とヒト疾病関連cDNA分離

分担研究者　菅野純夫　東京大学新領域創成科学研究科メディカルゲノム専攻
ゲノム制御医科学分野　教授

本年度は、カニクイザルの遺伝子のプロモーター部位をオリゴキャップ法による完全長cDNAの情報とヒト遺伝子のホモロジー情報をもとに、カニクイザル遺伝子のプロモータ領域（転写開始点上流1000塩基、下流200塩基）について、約380をクローン化した。これらにつき配列決定を進めるとともに、遺伝子発現の解析を行い、カニクイザルとヒトの比較を行った。

A．研究目的

悪性腫瘍・動脈硬化・糖尿病・精神病等疾病の理解とその解決ためには、システムとしての生体を遺伝子レベルで理解する事が必要である。このためにはヒトの全遺伝子が明らかにされ、その発現解析・機能解析が行われることが必要になる。

遺伝子の発現解析、機能解析には、mRNAの完全なコピーである完全長 cDNA が必要であり、遺伝子レベルで疾病を研究していくための欠くことの出来ない基盤となっている。特に、疾病の遺伝子レベルの研究を広範に展開するために、ヒトの全遺伝子の完全長 cDNA クローンを持つことは重要と考えられる。

本研究は、疾病の遺伝子レベルの研究基盤として、完全長 cDNA クローンを多数収集し、その情報をもとに、遺伝子発現の制御や、完全長 cDNA の機能解析に直結するような情報の収集を目指す。

B．研究方法

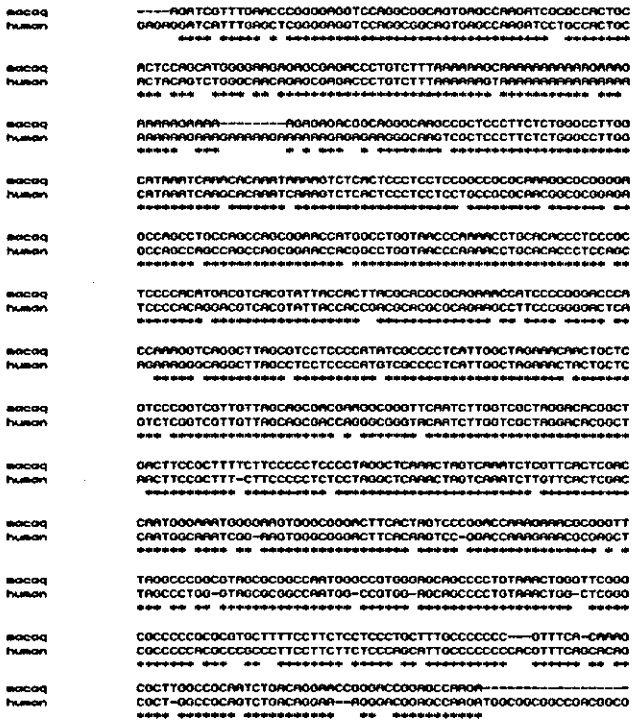
遺伝子調節領域はcDNA配列をもとに、主な調節配列が集まるとされる転写開始点近傍からゲノム上で上流約１０００bp、下流２００bpの領域のクローニングをおこなった。シークエンスによる配列の決定後、Clustal Wでアラインメントし、配列の比較をおこなった。さらに、両種間の遺伝子調節領域の配列の違いが遺伝子発現調節活性に関係するのか調べるべく、ルシフェラーゼ遺伝子をレポーターとするベクターを構築し、HEK293細胞にトランスフェクションし、４８時間後にルシフェラーゼアッセイをおこなった。

C．研究結果

本年度までにわれわれは、ヒトと他の霊長類との共通点、相違点を遺伝子レベルで明らかにすべく、カニクイザルの脳、心臓、腎臓、肝臓、精巣、皮膚などの組織から完全長cDNAライブラリーの作成をおこなった。さらに、約３万のワンパスシークエンスをおこない、得られたシークエンスデータを利用して、ヒトmRNAと翻訳開始コドン周辺配列の相同性、及び転写開始位置（TSS：Transcription start site）の分布様式について比較解析をおこない、①配列相同性は蛋白質コード領域(CDS)で97.8%、5’非翻訳領域（UTR）で94.7% ②TSSの分布様式は半数以上の遺伝子でよく保存されていたが、両種間で大きく異なっている遺伝

子もいくつかある、ということを示した。

図1　ppp2r1aのプロモーター比較

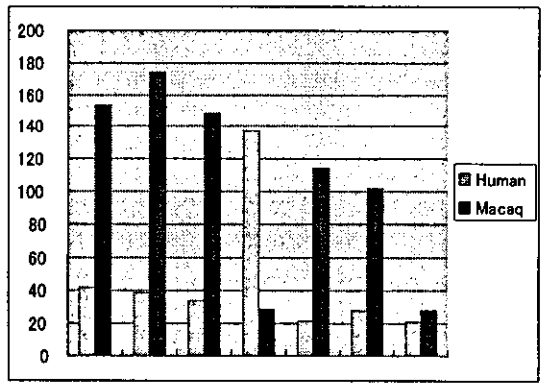


本年度はヒトとカニクイザルで転写制御様式に違いがある遺伝子の発見を目的として、プロモーター領域をクローン化して、発現を比較するために、カニクイザル遺伝子のプロモータ領域（転写開始点上流1000塩基、下流200塩基）について、約380をクローン化した。これらにつき配列決定を進めた。

ヒトとカニクイザル間でプロモーター領域およびUTRの配列相同性の低さとTSS分布様式に差が見られる遺伝子は、遺伝子発現制御に何らかの違いがある可能性が考えられた。遺伝子制御領域の配列相同性の極端に低い遺伝子としてppp2r1a（protein phosphatase2 regulatory subunitA alpha isoform）、mdh（malate dehydrogenase2, NAD）、TSS分布様式に差がなかった遺伝子としてalb（albmin）、plp1（ploteolipid protein 1）が得られた。すなわち、遺伝子調節領域の配列の比較から、ppp2r1aは配列相同性がCDSで98.7%、UTRが66.7%であるのに対し、遺伝子調節領域（896bp）の相同性は88.5%であった（図1）。また、albについては、CDS 96.7%、UTR 92.3%、遺伝子調節領域（1057bp）は93.4%であった。

これらの遺伝子について遺伝子調節領域の塩基配列の比較をおこなった。さらに、ルシフェラーゼアッセイにより、転写活性化能の実験的検証をおこなった。そのために、まず、ヒトとカニクイザルのppp2r1a遺伝子のプロモーターについて、連続欠失クローンを作製し、それらの活性について測定した（図2）。

図2　欠失クローンの転写活性



その結果、転写開始点上流200-300塩基の部分を欠失させると、カニクイザルでは活性が低下し、ヒトでは活性が上昇した。

D．考察

本年度は、カニクイザル遺伝子のプロモータ部分のクローン化を行い、さらに、そのプロモーター活性について、比較した。その結果コード領域に比べプロモーター配列における相同性が低い遺伝子が見出された。そのような遺伝子のプロモータ活性を詳細に検討すると、カニクイザルとヒトで相違が見出されることがわかった。こうし

た差がヒトとカニクイザルの差を生みだしている可能性もある。アルツハイマー病やエイズなどヒトで見られる疾患が他の霊長類で見られないことは良く知られており、こうした、微妙な発現の差の研究も疾患の治療に貢献する可能性がある。

E．結論

カニクイザル完全長cDNAライブラリーと5'端部分配列決定を組み合わせて、プロモーター領域を分離していくことは、遺伝子の発現制御研究の基盤を整備する上で有効であるばかりでなく、疾患の研究に重要な示唆を与えると考えられる。

F．研究発表

1．論文発表

1. Suzuki, Y., Yamashita, R., Sugano, S. and Nakai, K. DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. Nucleic Acids Res. 32: D78-81, 2004.

2. Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, Koyanagi KO, Barrero RA, Tamura T, Yamaguchi-Kabata Y, Tanino M, Yura K, Miyazaki S, Ikeo K, Homma K, Kasprzyk A, Nishikawa T, Hirakawa M, Thierry-Mieg J, Thierry-Mieg D, Ashurst J, Jia L, Nakao M, Thomas MA, Mulder N, Karavidopoulou Y, Jin L, Kim S, Yasuda T, Lenhard B, Eveno E, Suzuki Y, Yamasaki C, Takeda JI, Gough C, Hilton P, Fujii Y, Sakai H, Tanaka S, Amid C, Bellgard M, Bonaldo Md M, Bono H, Bromberg SK, Brookes AJ, Bruford E, Carninci P, Chelala C, Couillault C, Souza SJ, Debily MA, Devignes MD, Dubchak I, Endo T, Estreicher A, Eyras E, Fukami-Kobayashi K, R Gopinath G, Graudens E, Hahn Y, Han M, Han ZG, Hanada K, Hanaoka H, Harada E, Hashimoto K, Hinz U, Hirai M, Hishiki T, Hopkinson I, Imbeaud S, Inoko H, Kanapin A, Kaneko Y, Kasukawa T, Kelso J, Kersey P, Kikuno R, Kimura K, Korn B, Kuryshev V, Makalowska I, Makino T, Mano S, Mariage-Samson R, Mashima J, Matsuda H, Mewes HW, Minoshima S, Nagai K, Nagasaki H, Nagata N, Nigam R, Ogasawara O, Ohara O, Ohtsubo M, Okada N, Okido T, Oota S, Ota M, Ota T, Otsuki T, Piatier-Tonneau D, Poustka A, Ren SX, Saitou N, Sakai K, Sakamoto S, Sakate R, Schupp I, Servant F, Sherry S, Shiba R, Shimizu N, Shimoyama M, Simpson AJ, Soares B, Steward C, Suwa M, Suzuki M, Takahashi A, Tamiya G, Tanaka H, Taylor T, Terwilliger JD, Unneberg P, Veeramachaneni V, Watanabe S, Wilming L, Yasuda N, Yoo HS, Stodolsky M, Makalowski W, Go M, Nakai K, Takagi T, Kanehisa M, Sakaki Y, Quackenbush J, Okazaki Y, Hayashizaki Y, Hide W, Chakraborty R, Nishikawa K, Sugawara H, Tateno Y, Chen Z, Oishi M, Tonellato P, Apweiler R, Okubo K, Wagner L, Wiemann S, Strausberg RL, Isogai T, Auffray C, Nomura N, Gojobori T, Sugano S. Integrative Annotation of 21,037 Human Genes Validated by Full-Length cDNA Clones. PLoS Biol. 2004 Apr 20 [Epub ahead of print]

3. Suzuki Y, Yamashita R, Shirota M, Sakakibara Y, Chiba J, Mizushima-Sugano J, Nakai K, Sugano S. Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions. Genome Res. 14:1711-1718, 2004.

4. Bajic VB, Tan SL, Suzuki Y, Sugano S. Promoter prediction analysis on the whole human genome. Nat Biotechnol. 22: 1467-1473, 2004

5. Suzuki Y, Yamashita R, Shirota M, Sakakibara Y, Chiba J, Mizushima-Sugano J, Kel AE, Arakawa T, Carninci P, Kawai J, Hayashizaki Y, Takagi T, Nakai K, Sugano S. Large-scale collection and characterization of promoters of human and mouse genes. In Silico Biol. 4(3):0036 [Epub ahead of print] 2004

研究成果の刊行に関する一覧表

1) Imanishi, T. et al. (Hashimoto, K, Hirai, M. and Sugano, S.):
Integrative Annotation of 21,037 Human Genes Validated by Full-Length cDNA Clones.
PLoS Biology 2;6: 0856-0875, 2004.

2) Suzuki, Y., Yamashita, R., Sugano, S. and Nakai, K.:
DBTSS, DataBase of Transcriptional Start Sites: progress report 2004.
Nucleic Acids Res. 32: D78-81, 2004.

3) Suzuki Y, Yamashita R, Shirota M, Sakakibara Y, Chiba J, Mizushima-Sugano J, Nakai K, Sugano S.:
Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions.
Genome Research 14:1711-1718, 2004.

4) Bajic VB, Tan SL, Suzuki Y, Sugano S. :
Promoter prediction analysis on the whole human genome.
Nature Biotechnology 22: 1467-1473, 2004

5) Suzuki Y, Yamashita R, Shirota M, Sakakibara Y, Chiba J, Mizushima-Sugano J, Kel AE, Arakawa T, Carninci P, Kawai J, Hayashizaki Y, Takagi T, Nakai K, Sugano S.:
Large-scale collection and characterization of promoters of human and mouse genes.
In Silico Biology 4(3):0036 [Epub ahead of print] 2004

Integrative Annotation of 21,037 Human Genes Validated by Full-Length cDNA Clones

Tadashi Imanishi[1], Takeshi Itoh[1,2], Yutaka Suzuki[3,68], Claire O'Donovan[4], Satoshi Fukuchi[5], Kanako O. Koyanagi[6], Roberto A. Barrero[5], Takuro Tamura[7,8], Yumi Yamaguchi-Kabata[1], Motohiko Tanino[1,7], Kei Yura[9], Satoru Miyazaki[5], Kazuho Ikeo[5], Keiichi Homma[5], Arek Kasprzyk[4], Tetsuo Nishikawa[10,11], Mika Hirakawa[12], Jean Thierry-Mieg[13,14], Danielle Thierry-Mieg[13,14], Jennifer Ashurst[15], Libin Jia[16], Mitsuteru Nakao[3], Michael A. Thomas[17], Nicola Mulder[4], Youla Karavidopoulou[4], Lihua Jin[5], Sangsoo Kim[18], Tomohiro Yasuda[11], Boris Lenhard[19], Eric Eveno[20,21], Yoshiyuki Suzuki[5], Chisato Yamasaki[1], Jun-ichi Takeda[1], Craig Gough[1,7], Phillip Hilton[1,7], Yasuyuki Fujii[1,7], Hiroaki Sakai[1,7,22], Susumu Tanaka[1,7], Clara Amid[23], Matthew Bellgard[24], Maria de Fatima Bonaldo[25], Hidemasa Bono[26], Susan K. Bromberg[27], Anthony J. Brookes[19], Elspeth Bruford[28], Piero Carninci[29], Claude Chelala[20], Christine Couillault[20,21], Sandro J. de Souza[30], Marie-Anne Debily[20], Marie-Dominique Devignes[31], Inna Dubchak[32], Toshinori Endo[33], Anne Estreicher[34], Eduardo Eyras[15], Kaoru Fukami-Kobayashi[35], Gopal R. Gopinath[36], Esther Graudens[20,21], Yoonsoo Hahn[18], Michael Han[23], Ze-Guang Han[21,37], Kousuke Hanada[5], Hideki Hanaoka[1], Erimi Harada[1,7], Katsuyuki Hashimoto[38], Ursula Hinz[34], Momoki Hirai[39], Teruyoshi Hishiki[40], Ian Hopkinson[41,42], Sandrine Imbeaud[20,21], Hidetoshi Inoko[1,7,43], Alexander Kanapin[4], Yayoi Kaneko[1,7], Takeya Kasukawa[26], Janet Kelso[44], Paul Kersey[4], Reiko Kikuno[45], Kouichi Kimura[11], Bernhard Korn[46], Vladimir Kuryshev[47], Izabela Makalowska[48], Takashi Makino[5], Shuhei Mano[43], Regine Mariage-Samson[20], Jun Mashima[5], Hideo Matsuda[49], Hans-Werner Mewes[23], Shinsei Minoshima[50,52], Keiichi Nagai[11], Hideki Nagasaki[51], Naoki Nagata[1], Rajni Nigam[27], Osamu Ogasawara[3], Osamu Ohara[45], Masafumi Ohtsubo[52], Norihiro Okada[53], Toshihisa Okido[5], Satoshi Oota[35], Motonori Ota[54], Toshio Ota[22], Tetsuji Otsuki[55], Dominique Piatier-Tonneau[20], Annemarie Poustka[47], Shuang-Xi Ren[21,37], Naruya Saitou[56], Katsunaga Sakai[5], Shigetaka Sakamoto[5], Ryuichi Sakate[39], Ingo Schupp[47], Florence Servant[4], Stephen Sherry[13], Rie Shiba[1,7], Nobuyoshi Shimizu[52], Mary Shimoyama[27], Andrew J. Simpson[30], Bento Soares[25], Charles Steward[15], Makiko Suwa[51], Mami Suzuki[5], Aiko Takahashi[1,7], Gen Tamiya[1,7,43], Hiroshi Tanaka[33], Todd Taylor[57], Joseph D. Terwilliger[58], Per Unneberg[59], Vamsi Veeramachaneni[48], Shinya Watanabe[3], Laurens Wilming[15], Norikazu Yasuda[1,7], Hyang-Sook Yoo[18], Marvin Stodolsky[60], Wojciech Makalowski[48], Mitiko Go[61], Kenta Nakai[3], Toshihisa Takagi[3], Minoru Kanehisa[12], Yoshiyuki Sakaki[3,57], John Quackenbush[62], Yasushi Okazaki[26], Yoshihide Hayashizaki[26], Winston Hide[44], Ranajit Chakraborty[63], Ken Nishikawa[5], Hideaki Sugawara[5], Yoshio Tateno[5], Zhu Chen[21,37,64], Michio Oishi[45], Peter Tonellato[65], Rolf Apweiler[4], Kousaku Okubo[5,40], Lukas Wagner[13], Stefan Wiemann[47], Robert L. Strausberg[16], Takao Isogai[10,66], Charles Auffray[20,21], Nobuo Nomura[40], Takashi Gojobori[1,5,67*], Sumio Sugano[3,40,68]

1 Integrated Database Group, Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan, **2** Bioinformatics Laboratory, Genome Research Department, National Institute of Agrobiological Sciences, Ibaraki, Japan, **3** Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan, **4** EMBL Outstation—European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom, **5** Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Shizuoka, Japan, **6** Nara Institute of Science and Technology, Nara, Japan, **7** Integrated Database Group, Japan Biological Information Research Center, Japan Biological Informatics Consortium, Tokyo, Japan, **8** BITS Company, Shizuoka, Japan, **9** Quantum Bioinformatics Group, Center for Promotion of Computational Science and Engineering, Japan Atomic Energy Research Institute, Kyoto, Japan, **10** Reverse Proteomics Research Institute, Chiba, Japan, **11** Central Research Laboratory, Hitachi, Tokyo, Japan, **12** Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto, Japan, **13** National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America, **14** Centre National de la Recherche Scientifique (CNRS), Laboratoire de Physique Mathematique, Montpellier, France, **15** The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom, **16** National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America, **17** Department of Biological Sciences, Idaho State University, Pocatello, Idaho, United States of America, **18** Korea Research Institute of Bioscience and Biotechnology, Taejeon, Korea, **19** Center for Genomics and Bioinformatics, Karolinska Institutet, Stockholm, Sweden, **20** Genexpress—CNRS—Functional Genomics and Systemic Biology for Health, Villejuif Cedex, France, **21** Sino-French Laboratory in Life Sciences and Genomics, Shanghai, China, **22** Tokyo Research Laboratories, Kyowa Hakko Kogyo Company, Tokyo, Japan, **23** MIPS—Institute for Bioinformatics, GSF—National Research Center for Environment and Health, Neuherberg, Germany, **24** Centre for Bioinformatics and Biological Computing, School of Information Technology, Murdoch University, Murdoch, Western Australia, Australia, **25** Medical Education and Biomedical Research Facility, University of Iowa, Iowa City, Iowa, United States of America, **26** Genome Exploration Research Group, RIKEN Genomic Sciences Center, RIKEN Yokohama Institute, Kanagawa, Japan, **27** Medical College of Wisconsin, Milwaukee, Wisconsin, United States of America, **28** HUGO Gene Nomenclature Committee, University College London, London, United Kingdom, **29** Genome Science Laboratory, RIKEN, Saitama, Japan, **30** Ludwig Institute of Cancer Research, Sao Paulo, Brazil, **31** CNRS, Vandoeuvre les Nancy, France, **32** Lawrence Berkeley National Laboratory, Berkeley, California, United States of America, **33** Department of Bioinformatics, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan, **34** Swiss Institute of Bioinformatics, Geneva, Switzerland, **35** Bioresource Information Division, RIKEN BioResource Center, RIKEN Tsukuba Institute, Ibaraki, Japan, **36** Genome Knowledgebase, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America, **37** Chinese National Human Genome Center at Shanghai, Shanghai, China, **38** Division of Genetic Resources, National Institute of Infectious Diseases, Tokyo, Japan, **39** Graduate School of Frontier Sciences, Department of Integrated Biosciences, University of Tokyo, Chiba, Japan, **40** Functional Genomics Group, Biological Information Research Center, National Institute

of Advanced Industrial Science and Technology, Tokyo, Japan, **41** Department of Primary Care and Population Sciences, Royal Free University College Medical School, University College London, London, United Kingdom, **42** Clinical and Molecular Genetics Unit, The Institute of Child Health, London, United Kingdom, **43** Department of Genetic Information, Division of Molecular Life Science, School of Medicine, Tokai University, Kanagawa, Japan, **44** South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa, **45** Kazusa DNA Research Institute, Chiba, Japan, **46** RZPD Resource Center for Genome Research, Heidelberg, Germany, **47** Molecular Genome Analysis, German Cancer Research Center-DKFZ, Heidelberg, Germany, **48** Pennsylvania State University, University Park, Pennsylvania, United States of America, **49** Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, Osaka, Japan, **50** Medical Photobiology Department, Photon Medical Research Center, Hamamatsu University School of Medicine, Shizuoka, Japan, **51** Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan, **52** Department of Molecular Biology, Keio University School of Medicine, Tokyo, Japan, **53** Department of Biological Sciences, Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, Kanagawa, Japan, **54** Global Scientific Information and Computing Center, Tokyo Institute of Technology, Tokyo, Japan, **55** Molecular Biology Laboratory, Medicinal Research Laboratories, Taisho Pharmaceutical Company, Saitama, Japan, **56** Department of Population Genetics, National Institute of Genetics, Shizuoka, Japan, **57** Human Genome Research Group, Genomic Sciences Center, RIKEN Yokohama Institute, Kanagawa, Japan, **58** Columbia University and Columbia Genome Center, New York, New York, United States of America, **59** Department of Biotechnology, Royal Institute of Technology, Stockholm, Sweden, **60** Biology Division and Genome Task Group, Office of Biological and Environmental Research, United States Department of Energy, Washington, D.C., United States of America, **61** Faculty of Bio-Science, Nagahama Institute of Bio-Science and Technology, Shiga, Japan, **62** Institute for Genomic Research, Rockville, Maryland, United States of America, **63** Center for Genome Information, Department of Environmental Health, University of Cincinnati, Cincinnati, Ohio, United States of America, **64** State Key Laboratory of Medical Genomics, Shanghai Institute of Hematology, Rui-Jin Hospital, Shanghai Second Medical University, Shanghai, China, **65** PointOne Systems, Wauwatosa, Wisconsin, United States of America, **66** Graduate School of Life and Environmental Sciences, University of Tsukuba, Ibaraki, Japan, **67** Department of Genetics, Graduate University for Advanced Studies, Shizuoka, Japan, **68** Department of Medical Genome Sciences, Graduate School of Frontier Sciences, University of Tokyo, Tokyo, Japan

The human genome sequence defines our inherent biological potential; the realization of the biology encoded therein requires knowledge of the function of each gene. Currently, our knowledge in this area is still limited. Several lines of investigation have been used to elucidate the structure and function of the genes in the human genome. Even so, gene prediction remains a difficult task, as the varieties of transcripts of a gene may vary to a great extent. We thus performed an exhaustive integrative characterization of 41,118 full-length cDNAs that capture the gene transcripts as complete functional cassettes, providing an unequivocal report of structural and functional diversity at the gene level. Our international collaboration has validated 21,037 human gene candidates by analysis of high-quality full-length cDNA clones through curation using unified criteria. This led to the identification of 5,155 new gene candidates. It also manifested the most reliable way to control the quality of the cDNA clones. We have developed a human gene database, called the H-Invitational Database (H-InvDB; http://www.h-invitational.jp/). It provides the following: integrative annotation of human genes, description of gene structures, details of novel alternative splicing isoforms, non-protein-coding RNAs, functional domains, subcellular localizations, metabolic pathways, predictions of protein three-dimensional structure, mapping of known single nucleotide polymorphisms (SNPs), identification of polymorphic microsatellite repeats within human genes, and comparative results with mouse full-length cDNAs. The H-InvDB analysis has shown that up to 4% of the human genome sequence (National Center for Biotechnology Information build 34 assembly) may contain misassembled or missing regions. We found that 6.5% of the human gene candidates (1,377 loci) did not have a good protein-coding open reading frame, of which 296 loci are strong candidates for non-protein-coding RNA genes. In addition, among 72,027 uniquely mapped SNPs and insertions/deletions localized within human genes, 13,215 nonsynonymous SNPs, 315 nonsense SNPs, and 452 indels occurred in coding regions. Together with 25 polymorphic microsatellite repeats present in coding regions, they may alter protein structure, causing phenotypic effects or resulting in disease. The H-InvDB platform represents a substantial contribution to resources needed for the exploration of human biology and pathology.

Introduction

The draft sequences of the human, mouse, and rat genomes are already available (Lander et al. 2001; Marshall 2001; Venter et al. 2001; Waterston et al. 2002). The next challenge comes in the understanding of basic human molecular biology through interpretation of the human genome. To display biological data optimally we must first characterize the genome in terms of not only its structure but also function and diversity. It is of immediate interest to identify factors involved in the developmental process of organisms, non-protein-coding functional RNAs, the regulatory network of gene expression within tissues and its governance over states of health, and protein–gene and protein–protein interactions. In doing so, we must integrate this information in an easily accessible and intuitive format. The human genome may encode only 30,000 to 40,000 genes (Lander et al. 2001; Venter et al. 2001), suggesting that complex interde-

Received December 19, 2003; Accepted April 1, 2004; Published April 20, 2004
DOI: 10.1371/journal.pbio.0020162

Abbreviations: 3D, three-dimensional; AS, alternative splicing; CAI, codon adaptation index; dbSNP, Single Nucleotide Polymorphism Database; DDBJ, DNA Data Bank of Japan; EC, Enzyme Commission; EMBL, European Molecular Biology Laboratories; EST, expressed sequence tag; FANTOM, Functional Annotation of Mouse; FLcDNA, full-length cDNA; FLJ, Full-Length Long Japan; FTHFD, formyltetrahydrofolate dehydrogenase; GO, Gene Ontology; GTOP, Genomes TO Protein structures and functions database; H-Angel, Human Anatomic Gene Expression Library; H-Inv or H-Invitational, Human Full-Length cDNA Annotation Invitational; H-InvDB, H-Invitational Database; iAFLP, introduced amplified fragment length polymorphism; NCBI, National Center for Biotechnology Information; ncRNAs, non-protein-coding RNAs; OMIM, Online Mendelian Inheritance in Man; ORF, open reading frame; PDB, Protein Data Bank; RefSeq, Reference Sequence Collection; SMO, Similarity, Motif, and ORF; SNP, single nucleotide polymorphism

Academic Editor: Richard Roberts, New England Biolabs

*To whom correspondence should be addressed. E-mail: tgojobor@genes.nig.ac.jp

pendent gene regulation mechanisms exist to account for the complex gene networks that differentiate humans from lower-order organisms. In organisms with small genomes, it is relatively straightforward to use direct computational prediction based upon genomic sequence to identify most genes by their long open reading frames (ORFs). However, computational gene prediction from the genomic sequence of organisms with short exons and long introns can be somewhat error-prone (Ashburner 2000; Reese et al. 2000; Lander et al. 2001).

Previous efforts to catalogue the human transcriptome were based on expressed sequence tags (ESTs) used for the identification of new genes (Adams et al. 1991; Auffray et al. 1995; Houlgatte et al. 1995), chromosomal assignment of genes (Gieser and Swaroop 1992; Khan et al. 1992; Camargo et al. 2001), prediction of genes (Nomura et al. 1994), and assessment of gene expression (Okubo et al. 1992). Recently, Camargo et al. (2001) generated a large collection of ORF ESTs, and Saha et al. (2002) conducted a large-scale serial analysis of gene expression patterns to identify novel human genes. The availability of human full-length transcripts from many large-scale sequencing projects (Nomura et al. 1994; Nagase et al. 2001; Wiemann et al. 2001; Yudate 2001; Kikuno et al. 2002; Strausberg et al. 2002) has provided a unique opportunity for the comprehensive evaluation of the human transcriptome through the annotation of a variety of RNA transcripts. Protein-coding and non-protein-coding sequences, alternative splicing (AS) variants, and sense–antisense RNA pairs could all be functionally identified. We thus designed an international collaborative project to establish an integrative annotation database of 41,118 human full-length cDNAs (FLcDNAs). These cDNAs were collected from six high-throughput sequencing projects and evaluated at the first international jamboree, entitled the Human Full-length cDNA Annotation Invitational (H-Invitational or H-Inv) (Cyranoski 2002). This event was held in Tokyo, Japan, and took place from August 25 to September 3, 2002.

Efforts which have been made in the same area as the H-Inv annotation work include the Functional Annotation of Mouse (FANTOM) project (Kawai et al. 2001; Bono et al. 2002; Okazaki et al. 2002), Flybase (GOC 2001), and the RIKEN *Arabidopsis* full-length cDNA project (Seki et al. 2002). In our own project, great effort has been taken at all levels, not only in the annotation of the cDNAs but also in the way the data can be viewed and queried. These aspects, along with the applications of our research to disease research, distinguish our project from other similar projects.

This manuscript provides the first report by the H-Inv consortium, showing some of the discoveries made so far and introducing our new database of the human transcriptome. It is hoped that this will be the first in a long line of publications announcing discoveries made by the H-Inv consortium. Here we describe results from our integrative annotation in four major areas: mapping the transcriptome onto the human genome, functional annotation, polymorphism in the transcriptome, and evolution of the human transcriptome. We then introduce our new database of the human transcriptome, the H-Invitational Database (H-InvDB; http://www.h-invitational.jp), which stores all annotation results by the consortium. Free and unrestricted access to the H-Inv annotation work is available through the database. Finally, we summarize our most important findings thus far in the H-Inv project in Concluding Remarks.

Results/Discussion

Mapping the Transcriptome onto the Human Genome

Construction of the nonredundant human FLcDNA database. We present the first experimentally validated nonredundant transcriptome of human FLcDNAs produced by six high-throughput cDNA sequencing projects (Ota et al. 1997, 2004; Strausberg et al. 1999; Hu et al. 2000; Wiemann et al. 2001; Yudate 2001; Kikuno et al. 2002) as of July 15, 2002. The dataset consists of 41,118 cDNAs (H-Inv cDNAs) that were derived from 184 diverse cell types and tissues (see Dataset S1). The number of clones, the number of libraries, major tissue origins, methods, and URLs of cDNA clones for each cDNA project are summarized in Table 1. H-Inv cDNAs include 8,324 cDNAs recently identified by the Full-Length Long Japan (FLJ) project. The FLJ clones represent about half of the H-Inv cDNAs (Table 1). The policies for library selection and the results of initial analysis of the constituent projects were reported by the participants themselves: the Chinese National Human Genome Center (CHGC) (Hu et al. 2000), the Deutsches Krebsforschungszentrum (DKFZ/MIPS) (Wiemann et al. 2001), the Institute of Medical Science at the University of Tokyo (IMSUT) (Suzuki et al. 1997; Ota et al. 2004), the Kazusa cDNA sequence project of the Kazusa DNA Research Institute (KDRI) (Hirosawa et al. 1999; Nagase et al. 1999; Suyama et al. 1999; Kikuno et al. 2002), the Helix Research Institute (HRI) (Yudate et al. 2001), and the Mammalian Gene Collection (MGC) (Strausberg et al. 1999; Moonen et al. 2002), as well as FLJ mentioned earlier (Ota et al. 2004). The variation in tissue origins for library construction among these six groups resulted in rare occurrences of sequence redundancy among the collections. In a recent study, the FLJ project has described the complete sequencing and characterization of 21,243 human cDNAs (Ota et al. 2004). On the other hand, the H-Inv project characterized cDNAs from this project and six high-throughput cDNA producers by using a different suite of computational analysis techniques and an alternative system of functional annotation.

The 41,118 H-Inv cDNAs were mapped on to the human genome, and 40,140 were considered successfully aligned. The alignment criterion was that a cDNA was only aligned if it had both 95% identity and 90% length coverage against the genome (Figure 1). The mean identity of all the alignments between 40,140 mapped cDNAs and genomic sequences was 99.6 %, and the mean coverage against the genomic sequence was 99.6%. In some cases, terminal exons were aligned with low identity or low coverage. For example, 89% of internal exons have identity of 99.8% or higher, while only 78% and 50% of the first and last exons do, respectively. These alignments with low identity or low coverage seemed to be caused by the unsuccessful alignments of the repetitive sequences found in UTR regions and the misalignments of 3′ terminal poly-A sequences. Although better alignments could be obtained for these sequences by improving the mapping procedure, we concluded that the quality of the FLcDNAs was high overall.

Due to redundancy and AS within the human transcriptome, these 40,140 cDNAs were clustered to 20,190 loci

Table 1. Summary of cDNA Resources

cDNA Sequence Provider*	Number of cDNAs (Without Redundancy)	Number of Library Origins	Major Tissue Library Origins	Method	URL	Reference
CHGC	758 (754)	30	Adrenal gland, hypothalamus, CD34+ stem cell	Selecting FLcDNA clones from EST libraries	http://www.chgc.sh.cn/	Hu et al. 2000
DKFZ/MIPS	5,555 (5,521)	14	Testis, brain, lymph node	Selecting FLcDNA clones from 5′- and 3′- EST libraries	http://mips.gsf.de/projects/cdna	Wiemann et al. 2001
FLJ/HRI	8,066 (8,057)	46	Teratocarcinoma, placenta, whole embryo	Oligo-capping method and selection by one-pass sequences	http://www.hri.co.jp/HUNT/	Ota et al. 1997, 2004; Yudate et al. 2001
FLJ/IMSUT	12,585 (12,560)	81	Brain, testis, bone marrow	Oligo-capping method and selection by one-pass sequences	http://cdna.ims.u-tokyo.ac.jp/	Suzuki et al. 1997; Ota et al. 2004
FLJ/KDRI	348(342)	1	Spleen	Selection by one-pass sequences	http://www.kazusa.or.jp/NEDO/	Ota et al. 2004
KDRI	2,000 (2,000)	9	Brain	In vitro protein synthesis and selection by one-pass sequences	http://www.kazusa.or.jp/huge/	Hirosawa et al. 1997; Nagase et al. 1999; Suyama et al. 1999; Kikuno et al. 2002
MGC/NIH	11,806(11,414)	69	Placenta, lung, skin	Selecting gene candidates from 5′-EST libraries	http://mgc.nci.nih.gov/	Strausberg et al. 1999

*FLcDNA data were provided for H-Inv project by the FLJ project of NEDO (URL: http://www.nedo.go.jp/bio-e/) and six high-throughput cDNA clone producers Chinese National Human Genome Center (CHGC), the Deutsches Krebsforschungszentrum (DKFZ/MIPS), Helix Research Institute (HRI), the Institute of Medical Science in the University of Tokyo (IMSUT), the Kazusa DNA Research Institute (KDRI), and the Mammalian Gene Collection (MGC/NIH).
DOI: 10.1371/journal.pbio.0020162.t001

(H-Inv loci). For the remaining 978 unmapped cDNAs, we conducted cDNA-based clustering, which yielded 847 clusters. The clusters created had an average of 2.0 cDNAs per locus (Table 2). The average was only 1.2 for unmapped clusters, probably because many of these genes are encoded by heterochromatic regions of the human genome and show limited levels of gene expression. The gene density for each chromosome varied from 0.6 to 19.0 genes/Mb, with an average of 6.5 genes/Mb. This distribution of genes over the genome is far from random. This biased gene localization concurs with the gene density on chromosomes found in similar previous reports (Lander et al. 2001; Venter et al. 2001). This indicates that the sampled cDNAs are unbiased with respect to chromosomal location. Most cDNAs were mapped only at a single position on the human genome. However, 1,682 cDNAs could be mapped at multiple positions (with mean values of 98.2% identity and 98.1% coverage). The multiple matching may be caused by either recent gene duplication events or artificial duplication of the human genome caused by misassembled contigs. In our study we have selected only the "best" loci for the cDNAs (see Materials and Methods for details).

In total, 21,037 clusters (20,190 mapped and 847 unmapped) were identified and entered into the H-InvDB. We assigned H-Inv cluster IDs (e.g., HIX0000001) to the clusters and H-Inv cDNA IDs (e.g., HIT000000001) to all curated cDNAs. A representative sequence was selected from each cluster and used for further analyses and annotation.

Comparison of the mapped H-Inv cDNAs with other annotated datasets. In order to evaluate the H-Inv dataset, we compared all of the mapped H-Inv cDNAs with the Reference Sequence Collection (RefSeq) mRNA database (Pruitt and Maglott 2001) (Figure 2). The RefSeq mRNA database consists of two types of datasets. These are the curated mRNAs (accession prefix NM and NR) and the model mRNAs that are provided through automated processing of the genome annotation (accession prefix XM and XR).

From the comparison, we found that 5,155 (26%) of the H-Inv loci had no counterparts and were unique to the H-Inv. All of these 5,155 loci are candidates for new human genes, although non-protein-coding RNAs (ncRNAs) (25%), hypothetical proteins with ORFs less than 150 amino acids (55%), and singletons (91%) were enriched in this category. In fact, 1,340 of these H-Inv-unique loci were questionable and require validation by further experiments because they consist of only single exons, and the 3′ termini of these loci align with genomic poly-A sequences. This feature suggests internal poly-A priming although some occurrences might be bona fide genes. The most reliable set of newly identified human genes in our dataset is composed of 1,054 protein-

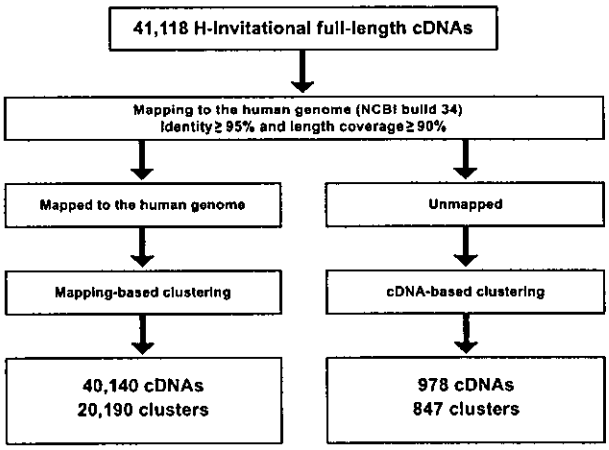


Figure 1. Procedure for Mapping and Clustering the H-Inv cDNAs

The cDNAs were mapped to the genome and clustered into loci. The remaining unmapped cDNAs were clustered based upon the grouping of significantly similar cDNAs.

DOI: 10.1371/journal.pbio.0020162.g001

coding and 179 non-protein-coding genes that have multiple exons. Therefore, at least 6.1% (1,233/20,190) of the H-Inv loci could be used to newly validate loci that the RefSeq datasets do not presently cover. These genes are possibly less expressed since the proportion of singletons (H-Inv loci consisting of a single H-Inv cDNA) was high (84%).

On the other hand, 78% (11,974/15,439) of the curated RefSeq mRNAs were covered by the H-Inv cDNAs. These figures suggest that further extensive sequencing of FLcDNA clones will be required in order to cover the entire human gene set. Nonetheless, this effort provides a systematic approach using the H-Inv cDNAs, even though a portion of the cDNAs have already been utilized in the RefSeq datasets.

It is noteworthy that H-Inv cDNAs overlapped 3,061 (17%) of RefSeq model mRNAs, supporting this proportion of the hypothetical RefSeq sequences. These newly confirmed 3,061 loci have a mean number of exons greater than RefSeq model mRNAs that were not confirmed, but smaller than RefSeq curated mRNAs. The overlap between H-Inv cDNAs and RefSeq model mRNAs was smaller than that between H-Inv cDNAs and RefSeq curated mRNAs. This suggests that the genes predicted from genome annotation may tend to be less expressed than RefSeq curated genes, or that some may be artifacts. All these results highlight the great importance of comprehensive collections of analyzed FLcDNAs for validat-

Table 2. The Clustering Results of Human FLcDNAs onto the Human Genome

Chromosome	Number of Loci	Number of cDNAs	Number of cDNAs/Locus	Number of Loci/Mb
1	1,998	4,057	2.0	8.1
2	1,408	2,791	2.0	5.8
3	1,224	2,455	2.0	6.1
4	809	1,527	1.9	4.2
5	920	1,851	2.0	5.1
6	1,027	1,912	1.9	6.0
7	1,008	1,994	2.0	6.4
8	761	1,448	1.9	5.2
9	817	1,630	2.0	6.0
10	863	1,705	2.0	6.4
11	1,116	2,245	2.0	8.3
12	1,014	2,071	2.0	7.7
13	394	743	1.9	3.5
14	626	1,363	2.2	5.9
15	693	1,415	2.0	6.9
16	865	1,851	2.1	9.6
17	1,110	2,245	2.0	13.6
18	334	593	1.8	4.4
19	1,210	2,378	2.0	19.0
20	536	1,124	2.1	8.4
21	197	379	1.9	4.2
22	480	985	2.1	9.7
X	646	1,173	1.8	4.2
Y	29	32	1.1	0.6
UN[a]	105	173	1.6	-
Unmapped	847	978	1.2	-
Total	21,037	41,118	2.0	-

[a]UN represents contigs that were not mapped onto any chromosome.

DOI: 10.1371/journal.pbio.0020162.t002

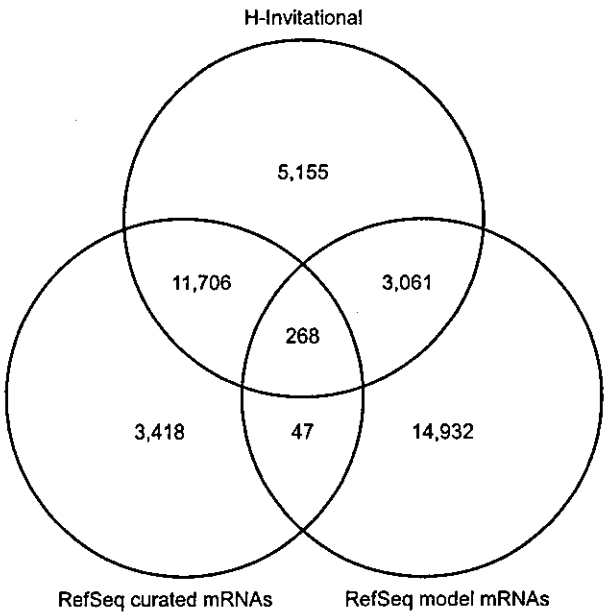


Figure 2. A Comparison of the Mapped H-Inv FLcDNAs and the RefSeq mRNAs

The mapped H-Inv cDNAs, the RefSeq curated mRNAs (accession prefixes NM and NR), and the RefSeq model mRNAs (accession prefixes XM and XR) provided by the genome annotation process were clustered based on the genome position. The numbers of loci that were identified by clustering are shown.

DOI: 10.1371/journal.pbio.0020162.g002

ing gene prediction from genome sequences. This may be especially true for higher organisms such as humans.

Incomplete parts of the human genome sequences. The existence of 978 unmapped cDNAs (847 clusters) suggests that the human genome sequence (National Center for Biotechnolgy Information [NCBI] build 34 assembly) is not yet complete. The evidence supporting this statement is twofold. First, most of those unmapped cDNAs could be partially mapped to the human genome. Using BLAST, 906 of the unmapped cDNAs (corresponding to 786 clusters) showed at least one sequence match to the human genome with a bit score higher than 100. Second, most of the cDNAs could be mapped unambiguously to the mouse genome sequences. A total of 907 unmapped cDNAs (779 clusters; 92%) could be mapped to the mouse genome with coverage of 90% or higher. If we adopted less stringent requirements, more cDNAs could be mapped to the mouse genome. The rest might be less conserved genes, genes in unfinished sections of the mouse genome, or genes that were lost in the mouse genome. Based on these observations, we conclude that the human genome sequence is not yet complete, leaving some portions to be sequenced or reassembled.

The proportion of the genome that is incomplete is estimated to be 3.7%–4.0%. The figure of 4.0% is based upon the proportion of H-Inv cDNA clusters that could not be mapped to the genome (847/21,037), while the 3.7% estimate is based on both H-Inv cDNAs and RefSeq sequences (only NMs). This statistic indicates that a minimum of one out of every 25–27 clusters appears to be unrepresented in the current human genome dataset, in its full form. Possible reasons for this include unsequenced regions on the human genome and regions where an error may have occurred during sequence assembly. If this is the case, this lends support to the use of cDNA mapping to facilitate the completion of whole genome sequences (Kent and Haussler 2001). For example, we can predict the arrangement of contigs based on the order of mapped exons. In addition we can use the sequences of unmapped exons to search for those clones that contain unsequenced parts of the genome. The mapping results of partially mapped cDNAs are thus quite useful.

Primary structure of genes on the human genome. Using the H-Inv cDNAs, the precise structures of many human genes could be identified based on the results of our cDNA mapping (Table S1). The median length of last exons (786 bp) was found to be longer than that of other exons, and the median length of first introns (3,152 bp) longer than that of other introns. These observed characteristics of human gene structures concur with the previous work using much smaller datasets (Hawkins 1988; Maroni 1996; Kriventseva and Gelfand 1999).

In the human genome, 50% of the sequence is occupied by repetitive elements (Lander et al. 2001). Repetitive elements were previously regarded by many as simply "junk" DNA. However, the contribution of these repetitive stretches to genome evolution has been suggested in recent works (Makalowski 2000; Deininger and Batzer 2002; Sorek et al. 2002; Lorenc and Makalowski 2003). The 21,037 loci of representative cDNAs were searched for repetitive elements using the RepeatMasker program. RepeatMasker indicated that 9,818 (47%) of the H-Inv cDNAs, including 5,442 coding hypothetical proteins, contained repetitive sequences. The existence of *Alu* repeats in 5% of human cDNAs was reported previously (Yulug et al. 1995). Our results revealed a significant number of repetitive sequences including *Alu* in the human transcriptome. Among them, 1,866 cDNAs overlapped repetitive sequences in their ORFs. Moreover, 554 of 1,866 cDNAs had repetitive sequences contained completely within their ORFs, including 81 cDNAs that were identical or similar to known proteins. This may indicate the involvement of repetitive elements in human transcriptome evolution, as suggested by the presence of Alu repeats in AS exons (Sorek et al. 2002) and the contribution to protein variability by repetitive elements in protein-coding regions (Makalowski 2000). We detected 2,254 and 5,427 cDNAs containing repetitive sequences in their 5′ UTR and 3′ UTR, respectively. The positioning of the repetitive elements suggests they play a regulatory role in the control of gene expression (Deininger and Batzer 2002) (see Table S1 or the H-InvDB for details).

AS transcripts. We wished to investigate the extent to which the functional diversity of the human proteome is affected by AS. In order to do this, we searched for potential AS isoforms in 7,874 loci that were supported by at least two H-Inv cDNAs. We examined whether or not these cDNAs represented mutually exclusive AS isoforms, using a combination of computational methods and human curation (see Materials and Methods). All AS isoforms that were supported independently by both methods were defined as the H-Inv AS dataset. Our analysis showed that 3,181 loci (40 % of the 7,874 loci) encoded 8,553 AS isoforms expressing a total of 18,612 AS exons. On average, 2.7 AS isoforms per locus were identified in these AS-containing loci. This figure represents

half of the AS isoforms predicted by another group (Lander et al. 2001). Our result highlights the degree to which full-length sequencing of redundant clones is necessary when characterizing the complete human transcriptome. The relative positions of AS exons on the loci varied: 4,383 isoforms comprising 1,538 loci were 5′ terminal AS variants; 5,678 isoforms comprising 1,979 loci were internal AS variants; and 2,524 isoforms comprising 921 loci were 3′ terminal AS variants.

The AS isoforms found in the H-Inv AS dataset have strikingly diverse functions. Motifs are found over a wide range of protein sequences. For certain types of subcellular targeting signals, such as signal peptides, position within the entire protein sequence appears crucial. A total of 3,020 (35 %) AS isoforms contained AS exons that overlapped protein-coding sequences. 1,660 out of 3,020 AS isoforms (55%) harbored AS exons that encoded functional motifs. Additionally, 1,475 loci encoded AS isoforms that had different subcellular localization signals, and 680 loci had AS isoforms that had different transmembrane domains. These results suggest marked functional differentiation between the varying isoforms. If this is the case, it would appear that AS contributes significantly to the functional diversity of the human proteome.

As the coverage of the human transcriptome by H-Inv cDNAs is incomplete, it would be misleading to conjecture that our dataset comprehensively includes all AS transcripts from every human gene. However, the current collection is a robust characterization of the existing functional diversity of the human proteome, and it represents a valuable resource of full-length clones for the characterization of experimentally determined AS isoforms.

In the cases where three-dimensional (3D) structures could be assigned to H-Inv cDNA protein products, we have examined the possible impact of AS rearrangements on the 3D structure. Our analysis was performed using the Genomes TO Protein structures and functions database (GTOP) (Kawabata et al. 2002). We found that some of the sequence regions in which internal exons vary between different isoforms contained regions encoding SCOP domains (Lo Conte et al. 2000). This discovery allowed us to perform a simple analysis of the structural effects of AS. Our analysis of the SCOP domain assignments revealed that the loci displaying AS are much more likely to contain class c (β–α–β units, α/β) SCOP domains than class d (segregated α and β regions, α+β) or class g (small) domains.

An example of exon differences between AS isoforms is presented in Figure 3. The structures shown are those of proteins in the Brookhaven Protein Data Bank (PDB) (Berman et al. 2000) to which the amino acid sequences of the corresponding AS isoforms are aligned. Segments of the AS isoform sequences that are not aligned with the corresponding 3D structure are shown in purple. Figure 3 demonstrates that exon differences resulting from AS sometimes give rise to significant alternations in 3D structure.

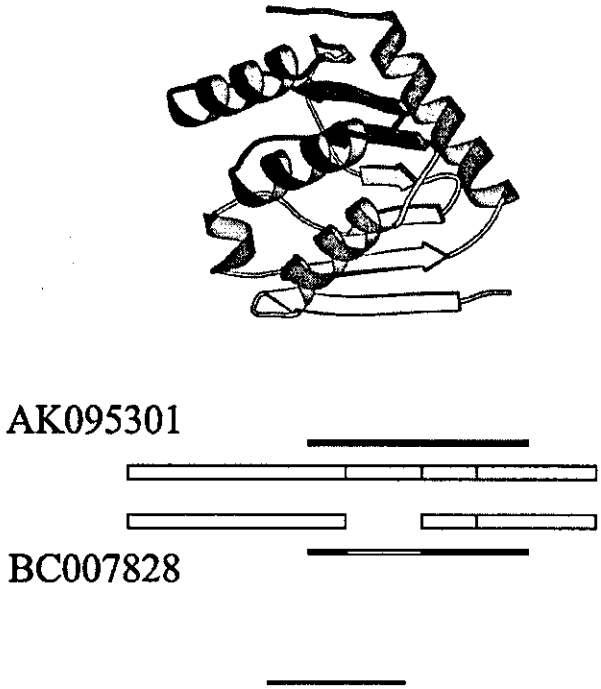


Figure 3. An Example of Different Structures Encoded by AS Variants

Exons are presented from the 5′ end, with those shared by AS variants aligned vertically. The AS variants, with accession numbers AK095301 and BC007828, are aligned to the SCOP domain d.136.1.1 and corresponding PDB structure 1byr. Helices and beta sheets are red and yellow, respectively. Green bars indicate regions aligned to the PDB structure, while open rectangles represent gaps in the alignments. AK095301 is aligned to the entire PDB structure shown, while BC007828 is lacking the alignment to the purple segment of the structure.

DOI: 10.1371/journal.pbio.0020162.g003

Functional Annotation

We predicted the ORFs of 41,118 H-Inv cDNA sequences using a computational approach (see Figure S1), of which 39,091 (95.1%) were protein coding and the remaining 2,027 (4.9%) were non-protein-coding. Since the structures and functions of protein products from AS isoforms are expected to be basically similar, we selected a "representative transcript" from each of the loci (see Figure S2). Then we identified 19,660 protein-coding and 1,377 non-protein-coding loci (Table 3). Human curation suggested that a total of 86 protein-coding transcripts should be deemed questionable transcripts. Once identified as dubious these sequences were excluded from further analysis. The remaining representatives from the 19,574 protein-coding loci were used to define a set of human proteins (H-Inv proteins). The tentative functions of the H-Inv proteins were predicted by computational methods. Following computational predictions was human curation.

After determination of the H-Inv proteins, we performed a standardized functional annotation as illustrated in Figure 4, during which we assigned the most suitable data source ID to each H-Inv protein based on the results of similarity search and InterProScan. We classified the 19,574 H-Inv proteins according to the levels of the sequence similarity. Using a system developed for the human cDNA annotation (see Figure S2), we classified the H-Inv proteins into five categories (Table 3). Three categories contain translated