I 第2章　絶対に必要な医学の基礎知識
第8節　その他のがん

長村文孝

はじめに

本項では，小児がんを除く，前節までに登場していない主要ながんについて簡潔にまとめる．特に，「症状」および「治療の副作用」について，特徴的なものは患者の心理状態に影響を与えることに留意されたい．

1．脳腫瘍

脳腫瘍は脳組織から発生する「原発性脳腫瘍」と，肺がんや乳がんなどの他のがんが脳に転移した「転移性脳腫瘍」とに分類され，原発性脳腫瘍は更に良性と悪性に分類される．良性腫瘍は周囲の正常組織と明確な境界を有し増大速度が一般に遅い．悪性腫瘍は逆に周囲との境界が不明瞭で，周辺あるいは脳の離れた部位に転移しやすく増大速度が速い．良性腫瘍には髄膜腫，下垂体腺腫，神経鞘腫などが含まれ，悪性脳腫瘍には膠芽腫，悪性星細胞腫，髄芽腫などが含まれる．それぞれの脳腫瘍の発生頻度は，年齢・性別により異なる．

脳腫瘍の症状は，大きく「局所症状」と「頭蓋内圧亢進症状」に分けられる．脳は感覚，運動などの中枢であり，それぞれの機能は脳の特定の部分に局在している．「局所症状」は，脳腫瘍によりある部位が制御している機能が障害を受けることによって発生する．脳では，ある機能を司る中枢の部位が決まっているため，「局所症状」として感覚障害，運動障害，視覚異常，言語障害などが発生する．言語障害は，舌・咽頭の運動障害により発音できないタイプ，言語自体を理解できないタイプ，話し方は流暢であるが内容が乏しく意味不明であるタイプなどに分かれる．このため，どのタイプなのかを把握して患者に接する必要がある．脳は頭蓋内という限られたスペースに存在するため，腫瘍が増大すると脳を圧迫し，脳圧が亢進する．このため，頭痛，嘔気，痙攣などが出現する．この症状を「頭蓋内圧亢進症状」という．この状態が長時間持続する場合には，失見当識や意識障害を呈することもある．下垂体は脳底にある小さな組織であるが，多くのホルモンの産生・調整に関わっている．下垂体の良性腫瘍はホルモンの過剰分泌を伴うことが多く，成長ホルモンが過剰分泌された場合には，子どもでは巨人症に，成人では末端肥大症を呈し，プロラクチン産生腫瘍では月経不順・無月経あるいは乳汁分泌を呈する．

脳腫瘍の診断には，CT，MRI，血管造影が用いられる．治療法は腫瘍摘出術，ガンマナイフを含む放射線療法あるいは化学療法が行われる．他の臓器と異なり，脳の場合には正常組織を切除することは障害の発生を意味するので，周囲への転移を考慮した十分な正常組織の切除を行うことができない．そのため，被膜を有さず，びまん性に周囲の正常組織に浸潤する悪性脳腫瘍を治癒的に切除することは通常困難である．化学療法は脳腫瘍の一部で施行されるが，治療効果は限定的である．良性腫瘍であっても，脳幹部などに発生

した場合には生命に関わるので，発生した場所および大きさ，近接する組織が重要となる．

2．泌尿器系のがん（腎がん，膀胱がん，前立腺がん）

1）腎がん

腎がんは小児に発生する「ウイルムス腫瘍」と，成人に発生する「腎細胞がん」に大別されるが，本項では腎細胞がんについて記載する．腎細胞がんは男性が女性よりも発生率が高く，喫煙や化学物質の曝露が危険因子とされている．腫瘍が小さい場合には無症状であり，検診などの超音波検査などで偶然発見されることが多い．腫瘍が大きくなると，血尿，背部不快感，背部痛をきたす．

診断は超音波検査，CTが有用である．治療は手術による切除が行われ，転移を有する場合には，インターフェロンやインターロイキン2を用いた免疫療法が行われている．インターフェロンは発熱などのインフルエンザに似た副作用が高頻度にみられるが，うつ状態あるいは自殺企図を呈することもあり，注意が必要である．近年，分子標的薬であるネクサバール®，スーテント®，トーリセル®が承認された．分子療法は一般に副作用が軽いと考えられる傾向にあるが，ネクサバール®では手足症候群（紅斑・水疱，皮膚の剥離），高血圧が，スーテント®では手足症候群，甲状腺機能障害が認められるように，特徴的な副作用が出現する．手足症候群では疼痛だけではなく，皮膚障害により日常の生活に著しく困難をきたす場合もある．

2）膀胱がん

膀胱がんは膀胱内面の移行上皮より発生する．発生は60歳以降で急激に増加し，男性が女性よりも発生率が高く，喫煙と化学物質の曝露が危険因子とされている．検診の尿検査で尿潜血陽性，あるいは超音波検査により発見される例もあるが，自覚症状としては肉眼的血尿と抗生剤に反応しない排尿時痛が多い．

治療は，がんが表在性であれば膀胱鏡を用いて，膀胱粘膜面に見られるがんを削除する経尿道的膀胱腫瘍切除術（transurethral resection of the bladder tumor：TUR-BT）が行われ，浸潤度が高い場合には膀胱の全摘が行われる．転移を有する場合には化学療法が行われ，再発予防として，あるいは複数の上皮内がんを有する場合には，膀胱内に結核のワクチンであるBCGや抗がん剤の注入も行われる．膀胱を全摘した場合には骨盤内郭清も行われるので，排尿障害，インポテンツなどの合併症に加え，膀胱の代替として尿をためておく部位（例：回腸導管造設術のストーマ）のケアが必要となる．

3）前立腺がん

前立腺がんは，腫瘍マーカーであるPSA（prostate specific antigen：前立腺特異抗原）を測定する健診によって無症状での発見も増えているが，排尿困難，頻尿，夜間多尿，残尿感など前立腺肥大と同じ症状により発見されることが多く，喫煙が危険因子とされている．検査は触診，超音波やCT，生検などが行われるが，骨転移が多いので骨シンチグラムも行われる．

前立腺がんは，高年齢であり，生検の結果，悪性度が低く，進行していない場合には，増殖速度が遅いため，年齢や臓器能を考慮して待機的に観察することもできる．手術による摘出でリンパ節郭清を行う場合には，膀胱がんの手術と同様に神経の損傷により排尿障害，インポテンツが起こる可能性が高い．放射線療法は身体の外から照射する外照射法と，小さな容器に放射線を発生するアイソトープを密封し前立腺に埋め込む密封小線源療法がある．後者は全身への影響が少ないが，熟達した医療機関が未だ十分に整備されていない状況である．このため，手術を選択するか放射線療法を選択するかで迷うケースが多い．前立腺がんは男性ホルモンによって増殖が促進されるので，精巣切除，抗男性ホルモン剤あるいは精巣で男性ホルモンが作られるのを阻害するLH-RHアナログ剤投与の内分泌療法が行われる．内分泌療法の問題点は，長期使用により効果が減弱すること，女性化乳房，乳房痛，急な発汗やのぼせ（ホットフラッシュ），脂肪の増加，性機能障害が挙げられる．外観が変わることがホルモン療法の継続を困難にすることも多い．骨転移においては，鎮痛剤などによる疼痛対策もQOL維持のためには重要である．

3．婦人科のがん（子宮頸がん，子宮体がん，卵巣がん）(図)

1）子宮頸がん

子宮のがんは，外子宮口近傍の頸部に発生する子宮頸がんと，子宮体部の内膜から発生する子宮体がんに大別される．

子宮頸がんは，90%以上の患者からヒューマン・パピローマ・ウイルス（human papillomavirus：HPV）が検出さるが，このウイルスは主に性行為により感染する．このため患者が「性病」の1つとして悩んだり，周囲から偏見を持たれる場合もあるので，カウンセリングの際には注意する必要がある．子宮頸がんは，子宮がん検診の細胞診により早期に発見されることが多く，この場合には凍結療法やレーザー照射，あるいは頸部の円

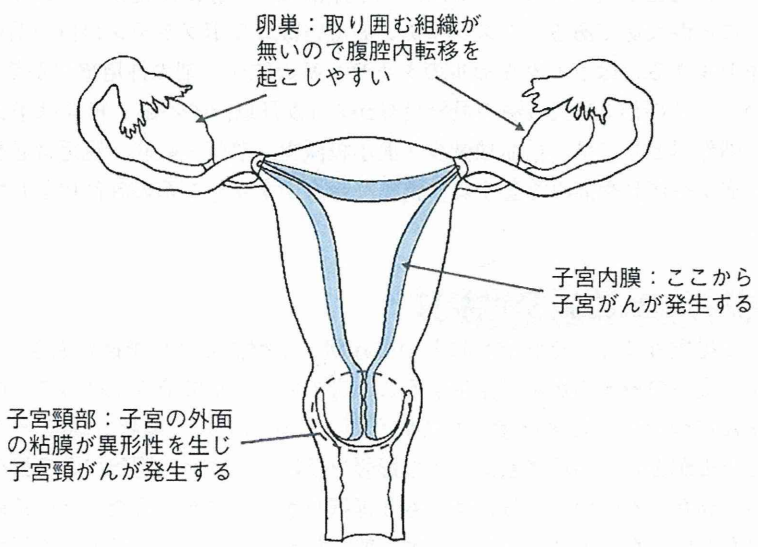

図　婦人科のがん

錐切除で治癒的に切除ができる．進展した場合には不正出血が自覚症状であることが多く，治療としては進展度に応じた手術切除，あるいは放射線療法が行われる．ヒューマン・パピローマ・ウイルスに対するワクチンがわが国でも認可され，発症の低下を期待して接種が拡まったが，意識消失，歩行障害などの副作用が報告されたため，接種が事実上行われなくなっている．

2)　子宮体がん

子宮体がんは，子宮内膜から発生するので子宮内膜がんとも呼ばれる．女性ホルモンの1つであるエストロゲンにより増殖が刺激されるタイプとそうでないものがある．高閉経年齢，未出産，肥満が主な危険因子である．不正出血，下腹部痛が主な症状で，診断には子宮の奥まで器材を挿入し組織を採取する必要がある．いわゆる「子宮がん検診」では子宮体がんを早期発見することはできない．治療は手術，放射線療法が主であり，化学療法は進行期に主に施行され，進行の度合いによって切除範囲が拡がっていく．早期であり，子宮を摘出したくない患者には，ホルモン療法も施行される場合がある．

3)　卵巣がん

卵巣がんは初期では自覚症状はなく，進行すると下腹部圧迫感，頻尿などの腫大による圧迫症状，転移による腹水貯留症状などの自覚症状が現れるので，受診して診断されることが多い．初期は卵巣の薄い皮膜内に留まっているが，増大し一度皮膜を破ると腹腔と遮る組織がないためすぐに腹腔内に転移を起こしやすい．しかし，早期発見を効率よく行える検診システムは確立されているとはいえない．診断のための検査は，超音波検査，CT，MRIと腫瘍マーカーであるCA125の測定が一般的に行われる．

治療法は外科手術と化学療法が主体である．早期は手術で治癒的切除が可能であるが，進行した場合には手術と化学療法を組み合わせた治療が行われる．卵巣がんは固形がんの中では比較的化学療法に反応するが，化学療法は標準療法が定まっており，これに則り施行することが大切である．シスプラチンあるいはカルボプラチンの白金系抗がん剤とパクリタキセルあるいはドセタキセルのタキサン系の抗がん剤の併用療法が第1選択療法とされている．この第1選択療法の最終投与から6カ月以内の不応・再発は予後不良とされている．副作用としては，白血球減少・血小板減少，嘔気・嘔吐，脱毛は必発であり，アレルギー症状や痺れや痛みを呈する末梢神経障害がタキサン系の副作用としてみられることがある．

サイコロジストへのメッセージ

がんの種類は多く，それぞれにおいて症状，治療法および予後が大きく異なり，1つの臓器から複数の種類のがんが発生することも多い．本項で取り上げることのできなかった他のがんについては他書を参考していただくしかないが，症状（QOLとも密接に関連する）と予後が患者の心理状態に大きな影響を及ぼす．また，選択できる標準療法があとどれくらい存在するのか，あるいは，もう選択肢がないのか，また，どの治療法を選択すべきなのかという点により，患者の心理状態に与えるインパクトは大きく異なることに留意が必要である．

I 第2章 絶対に必要な医学の知識
第12節 がん関連の臨床研究

長村文孝

はじめに

がんは，わが国において死亡原因の約1/3を占める．また，固形がんにおいては，局所の浸潤，あるいは転移により手術で十分に切除できない時等の化学療法の効果は，ほとんどの場合，生存期間の延長もしくは生活の質（quality of life：QOL）の向上であり，治癒は期待できない．近年，分子標的療法が出現し，旧来の殺細胞性の抗がん剤よりも副作用が軽減される傾向にあるが，がん以外の疾患に用いる薬剤と比較すると，副作用は一般的に高度である．このような状況を背景として，より有効な，あるいはより副作用の軽い治療方法を開発することが，「がん」の治療開発として急務である．そのため，新たな治療法開発・検証を目指した「臨床研究」あるいは「臨床試験」が盛んに行われている．

1．「臨床研究」，「臨床試験」とは

「臨床研究」とは，人を対象として，疾病の予防方法，診断方法および治療方法の改善，疾病原因および病態の理解並びに患者のQOLの向上を目的として実施される医学系研究である（厚生労働省「臨床研究に関する倫理指針」の「用語の定義」を改変）．これには，手術で摘出された検体，保存されている血清やDNAあるいは診療情報を用いた研究も含まれる．「臨床試験」は，「臨床研究」のうち，医薬品の投与あるいは医療機器を用いる等の被験者に対する介入行為を伴う研究であり，通常は研究計画を立て，実施計画書（プロトコール）の作成後に実施される（**図1**）．「治験」は，「臨床試験」のうち，医薬品

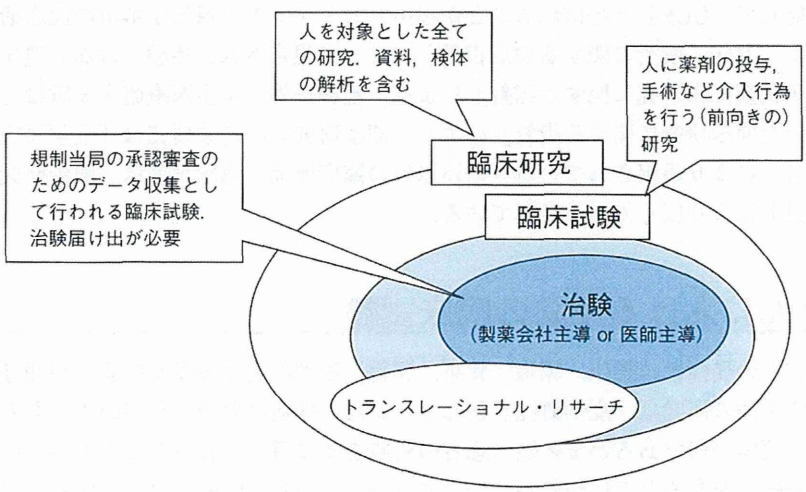

図1　臨床研究の分類

あるいは医療機器の厚生労働省等の規制当局からの製造・販売の認可を得るためのデータ収集を目的として行われる「臨床試験」を指し，試験開始に先立ち，わが国では医薬品医療機器総合機構（Pharmaceuticals and Medical Devices Agency：PMDA）を通じて厚生労働大臣宛に治験届けが提出され，PMDAで審査が行われる．「治験」は製薬企業等が主体のものと，医師が「自ら」実施する「医師主導治験」がある．後者は，対象患者数が少ないために製薬企業等が適応拡大等の治験を意図しない場合や，医師が研究者として自ら開発した治療法について承認申請を念頭に置き実施する場合に大別される．

「治験」ではない「臨床試験」は「自主臨床試験」あるいは「非治験臨床試験」等と呼ばれる．近年，なぜがんが発症するのか，あるいはがん細胞と正常細胞の違いは何か等の基礎研究が急速に発達し，これらの成果を基に臨床試験を行う「トランスレーショナルリサーチ（translational research：TR）」が盛んに行われるようになり，新たな治療法の開発として期待されている．TRは一般に早期の臨床試験に限定することが多く，「橋渡し研究」あるいは「探索型臨床研究」とも呼ばれる．抗がん剤の有効性，あるいは副作用の個人差等の臨床の結果を遺伝情報として研究することも多いが，このような臨床の結果から基礎に戻って行う研究もリーバースTRと呼称され，TRに含まれることがある．

2．臨床研究とガイドライン等

臨床研究の倫理的規範として，世界医師会が策定した「ヘルシンキ宣言」が知られている．研究参加のためのインフォームド・コンセント（informed consent：IC）の手続き，実施計画書の作成，倫理審査委員会での承認等，実施するための原則等から構成されている．他の法令や指針，あるいはガイドラインの憲法的存在となっていて，これを規範として臨床研究あるいは臨床試験の種類に対応して法規，指針等が定められている．

臨床研究・臨床試験には，その実施を定める法令やガイドラインが存在する．「治験」は薬事法とその関連省令に拠って規定されている．「治験」を規定する省令は「医薬品の臨床試験の実施の基準に関する省令」であり，英名が“Good Clinical Practice”であるので略して「GCP」と呼ばれることが多い．ヒトゲノムや遺伝子解析の臨床研究は「ヒトゲノム・遺伝子研究に関する倫理指針」によって規定され，治験ではない遺伝子治療は「遺伝子治療臨床研究に関する指針」により，治験以外の再生医療臨床試験は「ヒト幹細胞を用いた臨床研究に関する指針」により，観察研究等の疫学研究は「疫学研究に関する倫理指針」により規定されている．前述以外の臨床研究・臨床試験は「臨床研究に関する倫理指針」により広くカバーされている．

3．がんにおける予防の臨床試験

がんの発症には遺伝，環境，食事，薬物，物理的刺激等多くの要素が関与する．がんの発症予防の研究は「疫学研究」として，がんの発症の危険度を高める（あるいは低下させる）要因が何であるかを多数の患者の資料を基に解析されることが多い．疫学研究から，受動喫煙を含めた喫煙が肺がんだけではなく，口腔・咽頭・食道・膵臓・腎臓等多くのがんの発症リスクであることが判明した．このようなものには，B型肝炎ウイルス・C型肝

炎ウイルス感染による肝がんの発症，肥満における結腸がん，膵がん，閉経後乳がんなどが示されている．このように蓄積したデータの解析から危険因子の発見がなされるが，この危険因子が本当にがんの発症に関わっているのかを実証することは現実には非常に大変である．過去にはコーヒーが発がんの危険因子とされていたが，実際にはコーヒーの摂取は発がんのリスクではなかったことが判明したこともある．ヘリコバクター・ピロリ菌の場合では，除菌群と非除菌群の比較を前向き試験（prospective study）として試験計画を立案し，被験者を募集して実施する必要があるが，非常に多数の被験者と長期間の観察が必要であること，多額の研究資金が必要であること，リスクを参加者が認識していた場合には，それを回避する可能性があることから，臨床試験として実施することは困難な場合も多い．

がんでは，治療による副作用の発症予防の臨床試験も存在する．たとえば，化学療法後の口内炎予防のケア（例：ブラッシングの教育の成果）など高度な医療を用いなくとも実施できるテーマは多く存在し，また，心理面接の導入や面談等の工夫により，罹病によるうつ状態の改善や，前向きな心理状態への改善等を研究することも考えられる．

4．がん診断の臨床試験

診断の場合には，血液や尿等の検体の解析を基にするものと，画像診断等の医療機器に関するものとに大別される．尿や便を使用する場合は体への侵襲はなく，採血の場合には侵襲性は低く，倫理的判断をあまり要しないことが多い．しかし，被曝等の副作用の可能性のある医療機器を使用する場合や，ラジオ・アイソトープを体内に注射し放射線の被曝を受ける場合等では，安全性・有益性の十分な検討と被験者へ適切なICが必須となる．

5．がんの治療法の臨床試験

治療においても，過去のデータを基に推察する「臨床研究」と新たに介入行為を行う「臨床試験」に大別されるが，一般的には後者を指す．また，医薬品等の投与と手術等の手技においては「臨床試験」の概念がやや異なる．医薬品の臨床試験は第一相試験から第三相試験まで分類され，新薬開発では第一相試験から行われる．第一相試験は，動物実験等による非臨床試験での結果を基に，人への投与量，投与経路，投与スケジュール等を十分に検討した後に，人に初めて投与する段階である．抗がん剤以外の医薬品では，安全性，薬効の有無および薬理学的な情報の収集を目的として，健常人あるいは軽症の患者を対象として行われる．一方，抗がん剤の第一相試験は対象が異なり通常生存期間を延長する標準療法の無い段階の患者，すなわち標準療法の無くなった段階の患者を対象として行う．また，投与量を段階的に増加する用量漸増試験として行われ，人に投与できる最大耐用量（maximum tolerated dose：MTD）を決定する試験デザインであることが多い．これは，抗がん剤は，① 副作用が強い場合や，催奇性や2次発がんの危険性により健常人あるいは長期生存が見込まれる患者に投与することは倫理的に許容できない場合が多いこと，② 一般に抗がん剤は有効性と安全性を考慮した至適投与量の幅が狭く，有効性の低い低投与群あるいは毒性の強い高投与群の被験者が存在し，治療としての恩恵にあずかる

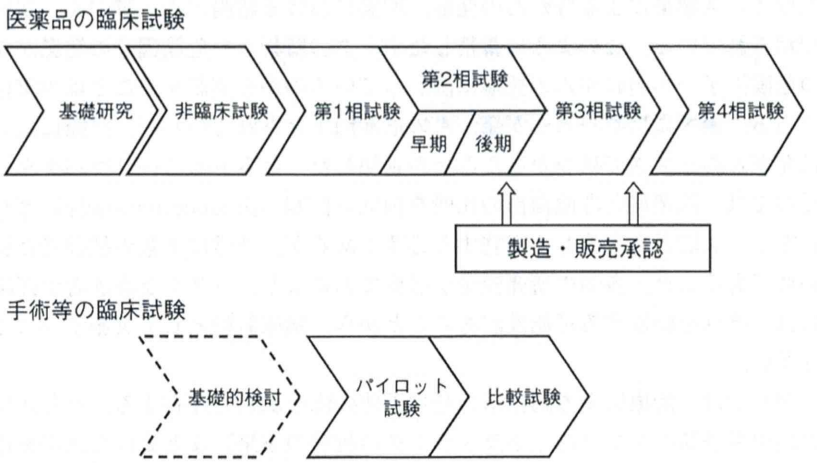

図2　臨床試験の流れ

ことのできる可能性のない被験者，あるいは毒性が強く生命の危機が生じる可能性のある被験者も存在することによる．これらを背景として，第一相試験は，被験者に有効性をもたらす可能性は低い．しかしながら，治癒の可能性を信じている患者も多いことに留意する必要がある．第二相試験は，第一相試験の結果を基に投与量・投与スケジュールを決定した後に，数十人程度の患者に投与し，有効性・安全性のある程度の情報を得ることを目的とする．1つの薬剤を単一投与量で最初の第二相試験を行うことが多いが，投与量の比較あるいは標準薬との比較試験として行われることある．また，「第二相試験」として行われる場合と，「早期第二相試験」と「後期第二相試験」とに分けて行われる場合がある．第三相試験は，標準療法との比較による有効性・安全性の検証を目的として数百人規模で実施される．迅速な第三相試験の実施と承認を目的として，最近は国際共同治験も増えている．第四相試験は市販後に情報を収集するために行われる．手術などの手技の臨床試験では，試験の分類が細分化されておらず，新たな手法を探索的に検討するパイロット試験と標準手技との比較試験に大別される（図2）．

臨床試験では，試験の目的を具体的数値で評価するためにエンドポイントを設定する．がんにおいては，最終的な目的は生存期間の延長あるいはQOLの向上であるが，これらを評価するためには多くの被験者と長い期間を必要とする．したがって，早期の試験では，これらと因果関係を有すると考えられる奏功率（腫瘍が縮小した被験者の割合）をサロゲート（代替）・エンドポイントとして用いることが多い．

サイコロジストへのメッセージ

臨床研究，特に臨床試験においては，被験者に対する十分な説明と適切な同意の取得が必須である．そのためには試験の目的，方法の理解が必要であり，更にその背景の理解として，対象とするがんの知識も必要となる．臨床試験が対象とするがんの種類，段階，予後，比較試験の場合には，その対象となる治療法等により患者の不安，選択への迷い，あるいは過度の期待が懸念されるので十分な知識をもって接することが重要である．

ARTICLE

doi:10.1038/nature13182

A promoter-level mammalian expression atlas

The FANTOM Consortium and the RIKEN PMI and CLST (DGT)*

Regulated transcription controls the diversity, developmental pathways and spatial organization of the hundreds of cell types that make up a mammal. Using single-molecule cDNA sequencing, we mapped transcription start sites (TSSs) and their usage in human and mouse primary cells, cell lines and tissues to produce a comprehensive overview of mammalian gene expression across the human body. We find that few genes are truly 'housekeeping', whereas many mammalian promoters are composite entities composed of several closely separated TSSs, with independent cell-type-specific expression profiles. TSSs specific to different cell types evolve at different rates, whereas promoters of broadly expressed genes are the most conserved. Promoter-based expression analysis reveals key transcription factors defining cell states and links them to binding-site motifs. The functions of identified novel transcripts can be predicted by coexpression and sample ontology enrichment analyses. The functional annotation of the mammalian genome 5 (FANTOM5) project provides comprehensive expression profiles and functional annotation of mammalian cell-type-specific transcriptomes with wide applications in biomedical research.

The mammalian genome encodes the instructions to specify development from the zygote through gastrulation, implantation and generation of the full set of organs necessary to become an adult, to respond to environmental influences, and eventually to reproduce. Although the genome information is the same in almost all cells of an individual, at least 400 distinct cell types[1] have their own regulatory repertoire of active and inactive genes. Each cell type responds acutely to alterations in its environment with changes in gene expression, and interacts with other cells to generate complex activities such as movement, vision, memory and immune response.

Identities of cell types are determined by transcriptional cascades that start initially in the fertilised egg. In each cell lineage, specific sets of transcription factors are induced or repressed. These factors together provide proximal and distal regulatory inputs that are integrated at transcription start sites (TSSs) to control the transcription of target genes. Most genes have more than one TSS, and the regulatory inputs that determine TSS choice and activity are diverse and complex (reviewed in ref. 2).

Unbiased annotation of the regulation, expression and function of mammalian genes requires systematic sampling of the distinct mammalian cell types and methods that can identify the set of TSSs and transcription factors that regulate their utilization. To this end, the FANTOM5 project has performed cap analysis of gene expression (CAGE)[3] across 975 human and 399 mouse samples, including primary cells, tissues and cancer cell lines, using single-molecule sequencing[3] (Fig. 1; see the full sample list in Supplementary Table 1).

CAGE libraries were sequenced to a median depth of 4 million mapped tags per sample (Supplementary Methods) to produce a unique gene expression profile, focused specifically on promoter utilization. CAGE has advantages over RNA-seq or microarrays for this purpose, because it permits separate analysis of multiple promoters linked to the same gene[13]. Moreover, we show in an accompanying manuscript[4] that the data can be used to locate active enhancers, and to provide numerous insights into cell-type-specific transcriptional regulatory networks (see the FANTOM5 website http://fantom.gsc.riken.jp/5). The data extend and complement the recently published ENCODE[5] data, and microarray-based gene expression atlases[6] to provide a major resource for functional genome annotation and for understanding the transcriptional networks underpinning mammalian cellular differentiation.

The FANTOM5 promoter atlas

Single molecule CAGE profiles were generated across a collection of 573 human primary cell samples (~ 3 donors for most cell types) and 128 mouse primary cell samples, covering most mammalian cell steady states. This data set is complemented with profiles of 250 different cancer cell lines (all available through public repositories and representing 154 distinct cancer subtypes), 152 human post-mortem tissues and 271 mouse developmental tissue samples (Fig. 1a; see the full sample list in Supplementary Table 1). To facilitate data mining all samples were annotated using structured ontologies (Cell Ontology[7], Uberon[8], Disease Ontology[9]). The results of all analyses are summarized in the FANTOM5 online resource (http://fantom.gsc.riken.jp/5). We also developed two specialized tools for exploration of the data. ZENBU, based on the genome browser concept, allows users to interactively explore the relationship between genomic distribution of CAGE tags and expression profiles[10]. SSTAR, an interconnected semantic tool, allows users to explore the relationships between genes, promoters, samples, transcription factors, transcription factor binding sites and coexpressed sets of promoters. These and other ways to access the data are described in more detail in Supplementary Note 1.

CAGE peak identification and thresholding

To identify CAGE peaks across the genome we developed decomposition-based peak identification (DPI; described in Supplementary Methods; Extended Data Fig. 1). This method first clusters CAGE tags based on proximity. For clusters wider than 49 base pairs (bp) it attempts to decompose the signal into non-overlapping sub-regions with different expression profiles using independent component analysis[11]. Sample- and genome-wide, DPI identified 3,492,729 peaks in human and 2,088,255 peaks in mouse. To minimize the fraction of peaks[3] that map to internal exons (which could exist due to post-transcriptional cleavage and recapping of RNAs[12]), and enrich for TSSs, we applied tag evidence thresholds

*Lists of participants and their affiliations appear at the end of the paper.

Figure 1 | Promoter discovery and definition in FANTOM5. **a**, Samples profiled in FANTOM5. **b**, Reproducible cell-type-specific CAGE patterns observed for the 266 base CpG island associated *B4GALT1* locus transcription initiation region hg19:chr9:33167138..33167403. CAGE profiles for $CD4^+$ T cells (blue), $CD14^+$ monocytes (gold), aortic smooth muscle cells (green) and the adrenal cortex adenocarcinoma cell line SW-13 (red) are shown. A combined pooled profile showing TSS distribution across the entire human collection is shown in black. Values on the *y* axis correspond to maximum normalized TPM for a single base in each track. **c**, Decomposition-based peak identification (DPI) finds 6 differentially used peaks within this composite transcription initiation region (note: peaks are labelled from p1@B4GALT1 with most tag support through to p7@B4GALT1 with the least tag support; p4@B4GALT1 is not shown and is in the 3′ UTR of the locus at position hg19::chr9:33111241..33111254−). Note in particular one large broad region on the left used in all samples and a sharp peak to the right, preferentially used in the aortic smooth muscle cells. **d**, Venn diagram showing DPI defined peaks expressed at ≥ 10 TPM in primary cells (red), tissues (blue) and cell lines (green). **e**, Fraction of unannotated peaks observed in subsets of **d**. P, primary cells, T, tissues, C, cell lines, PT, TC, PC and PTC correspond to peaks found in multiple sample types, for example, PT, found in primary cells and tissue samples.

to define robust and permissive subsets (described in more detail in Supplementary Methods and summarized in Table 1). Specifically the robust threshold, which is used for most of the analyses presented here, enriched for peaks at known 5′ ends compared to known internal exons by twofold (that is, two-thirds of the peaks hitting known full-length transcript models hit the 5′ end). A flow diagram showing the relationship between samples, peaks, thresholding and subsets used in each analysis is provided in the Supplementary Figure 1. Supporting evidence that the peaks are genuine TSSs, based upon support from expressed sequence tags (ESTs), histone H3 lysine 4 trimethylation (H3K4Me3) marks and DNase hypersensitive sites is provided in Supplementary Note 2.

Figure 1b illustrates the 266 bp spanning transcription initiation region of *B4GALT1*, where 6 independent robust peaks were identified by DPI, each with a unique regulatory pattern (Fig. 1c). A total of 58% of human and 56% of mouse robust peaks occur in such composite transcription initiation regions, defined as clusters of robust peaks within 100 bases of each other. More than half of these contain peaks with statistically significant differences in expression profiles (63% of human and 54% of mouse composite transcription initiation regions; likelihood ratio test, false discovery rate (FDR) <1%, Extended Data Fig. 1d). Supplementary Tables 2 and 3 summarize public domain EST evidence that these independent peaks contained within composite transcription initiation regions give rise to long RNAs.

Known gene coverage in FANTOM5

To provide annotation of the CAGE peaks, the distance between individual peaks and the 5′ ends of known full-length transcripts was determined and then peaks within 500 bases of the 5′ end of known transcript models were assigned to that gene (see Supplementary Methods, Table 1). To provide names for each TSS region, peaks identified at the permissive threshold were ranked by the total number of tags supporting each and then sequentially numbered (for example, p1@GFAP corresponds to the promoter of *GFAP* which has the highest tag support). From these annotations, TSS for 91% of human protein coding genes (as defined by the HUGO Gene Nomenclature Committee) were supported by robust CAGE peaks, and 94% at the permissive threshold (Supplementary Note 3). The atlas also detected signals from the promoters of short RNA primary transcripts, and long non-coding RNAs. In comparison to the previous FANTOM3 and 4 projects, FANTOM5 measured expression at an additional 4,721 human and 5,127 mouse RefSeq genes. The inclusion of primary cells, cell lines and tissues in the atlas provided greater coverage than any of the sample types alone (Fig. 1d) and the primary cell samples in particular were a rich source of unannotated peaks (Fig. 1e).

Mammalian promoter architectures

Mammalian promoters can be classified as broad or sharp types, based upon local spread of TSSs along the genome[13]. The FANTOM5 data