RNAシーケンスを始めよう

第4回
アリル特異的発現解析から融合遺伝子探索、*de novo* アセンブルまで

2011年12月

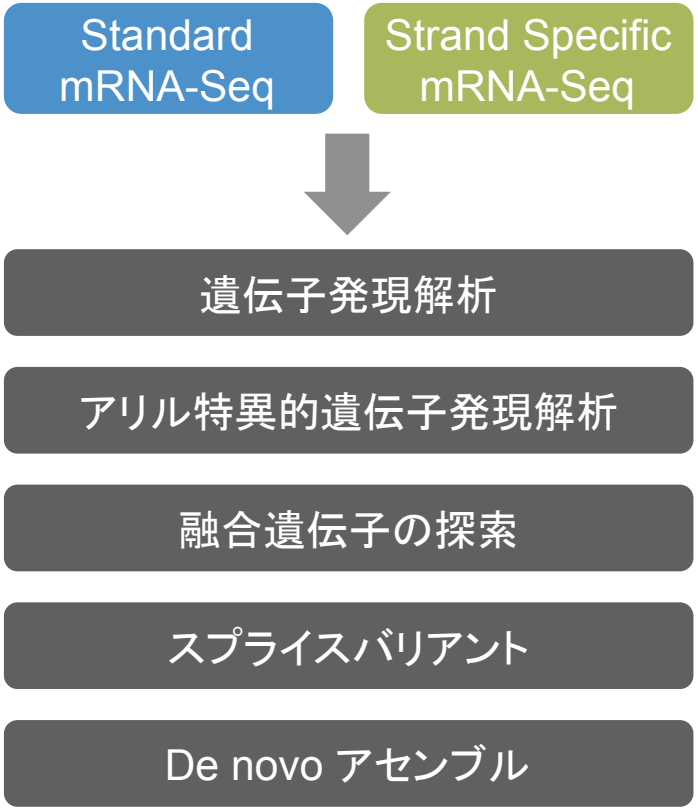
イルミナ株式会社
マーケティング部
鈴木 健介

illumina®

今日の内容

- サンプル調製キットのまとめ
  - Standard mRNA-Seq
  - Strand Specific mRNA-Seq

- 応用事例
  - 遺伝子発現解析アリル特異的遺伝子発現
  - 融合遺伝子探索
  - スプライスバリアント
  - De novo アプリケーション



今日の内容

- サンプル調製キットのまとめ
  - Standard mRNA-Seq
  - Strand Specific mRNA-Seq
- 応用事例
  - 遺伝子発現解析
  - アリル特異的遺伝子発現
  - 融合遺伝子探索
  - スプライスバリアント
  - De novo アプリケーション

mRNA-Seqには主に2つのプロトコールが存在

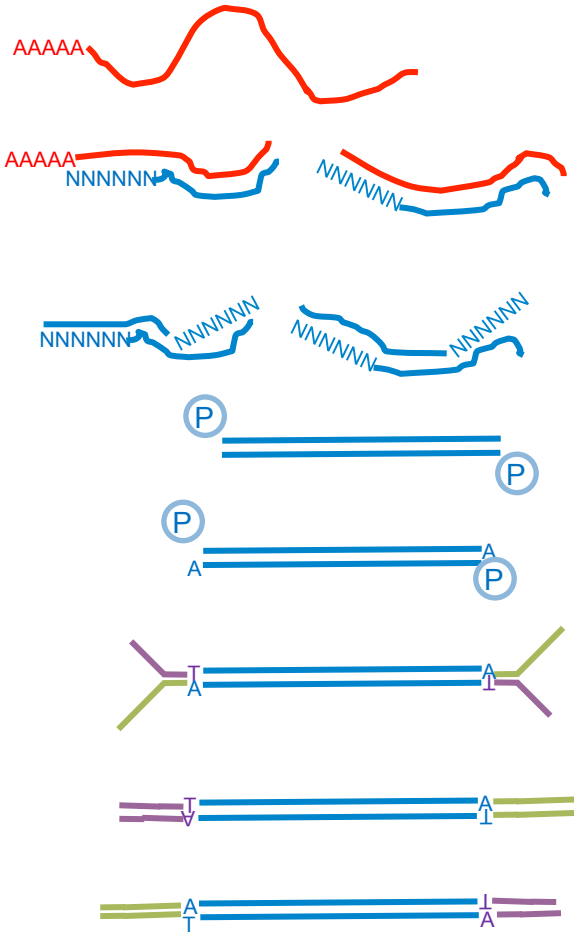
illumina®　　epicentre® an Illumina® company

	Standard	Strand Specific
キット	• TruSeq RNA Sample Prep Kit	• Epicentre ScriptSeq™ Sample Prep Kit
スタート材料	• Total RNA • 1ug	• Poly A 、あるいは rRNA 除去処理した RNA • 50-250 ng
ワークフロー	• Poly A およびランダムプライマーを使い、2本鎖cDNA を合成	• ランダムプライマーとタグ配列を使い、2本鎖 cDNA を合成 • タグ配列でストランドを認識
利点	• 標準的な遺伝子発現解析手法 • サンプルあたりのコストが安価	• 遺伝子発現に加えてストランド情報を取得 • バクテリア、FFPEにも応用可能

Standard mRNA-Seq のワークフロー

TruSeq RNA Sample Prep Kit

illumina®



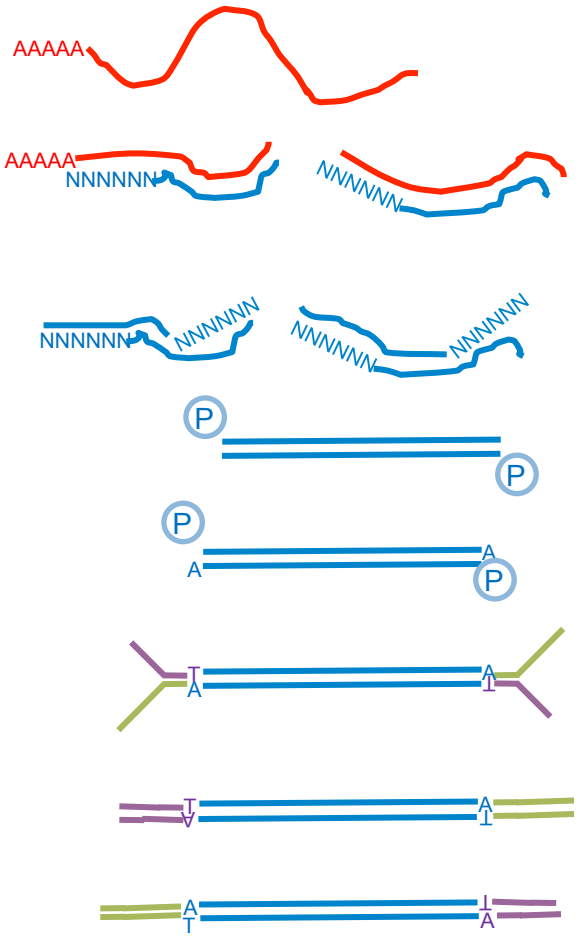
illumina®

Standard mRNA-Seq のワークフロー

TruSeq RNA Sample Prep Kit

1日

- Total RNA から poly A を使い mRNA を回収
- RNA の断片化
- ランダムプライマーを用いた cDNA 合成
- 2本鎖 cDNA 合成
- 末端修復、リン酸化、Aテイル付加
- インデックス付アダプターのライゲーション
- ビーズによるサイズ選択
- 15サイクルPCR



illumina®

スタート材料が微量の場合に使用できるキット

- Clonetech SMARTer Ultra Low RNA Kit
  - 日本国内ではタカラ株式会社で販売およびサポート

1日

Total RNA* (0.1～10ng)スタート

1本鎖 full length cDNA 合成＆精製

LD-PCR による2本鎖 cDNA の増幅合成

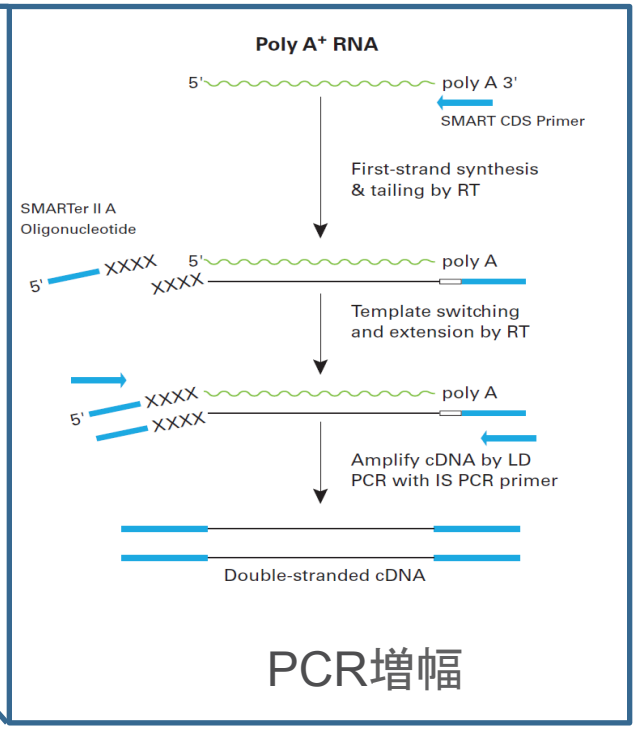
増幅 cDNA の精製 & 検証

増幅 cDNA の断片化（Covaris）

cDNA QC

1日

イルミナ DNA Paired End Sampe Prep Kit
（改変版）



illumina

スタート材料が微量の場合に使用できるキット

- Clonetech SMARTer Ultra Low RNA Kit
  - 日本国内ではタカラ株式会社で販売およびサポート

1日

Total RNA* (0.1～10ng)スタート

1本鎖 full length cDNA 合成＆精製

LD-PCR による2本鎖 cDNA の増幅合成

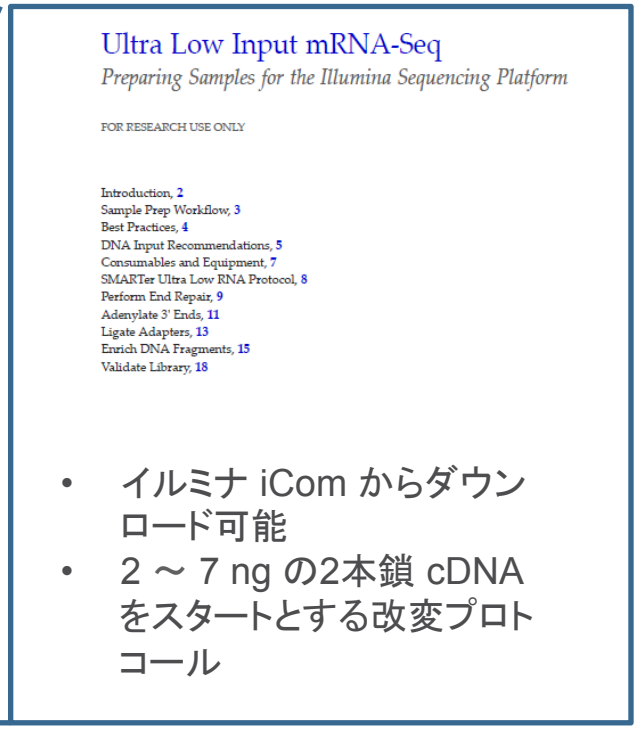
増幅 cDNA の精製 & 検証

増幅 cDNA の断片化（Covaris）

cDNA QC

1日

イルミナ DNA Paired End Sampe Prep Kit
（改変版）



Ultra Low Input mRNA-Seq

Preparing Samples for the Illumina Sequencing Platform

FOR RESEARCH USE ONLY

- イルミナ iCom からダウンロード可能
- 2 ～ 7 ng の2本鎖 cDNA をスタートとする改変プロトコール

illumina

サンプル調製キットのまとめ; Standard mRNA-Seq

通常プロトコール

カタログ番号	製品名	キット価格	サンプルあたりの価格	問い合せ
RS-122-2001	TruSeq RNA Sample Prep Kit v2 - Set A (48 samples)	503,000円	10,479円	イルミナ
RS-122-2002	TruSeq RNA Sample Prep Kit v2 - Set B (48 samples)	503,000円	10,479円	イルミナ

- 各キット12種類のインデックスを含みます（最大24種類）

微量DNAスタートの場合

カタログ番号	製品名	キット価格	サンプルあたりの価格	問い合せ
634935	SMARTer™ Ultra Low RNA Kit for Illumina ® Sequencing (10 Reactions)	228,000円	22,800円	タカラバイオ
PE-102-1001	Paired End Sample Prep Kit (10 Samples)	540,000円	54,000円	イルミナ

- タカラバイオ株式会社　問い合わせ先 077-543-6116

mRNA-Seqには主に2つのプロトコールが存在

illumina®

epicentre®
an Illumina® company

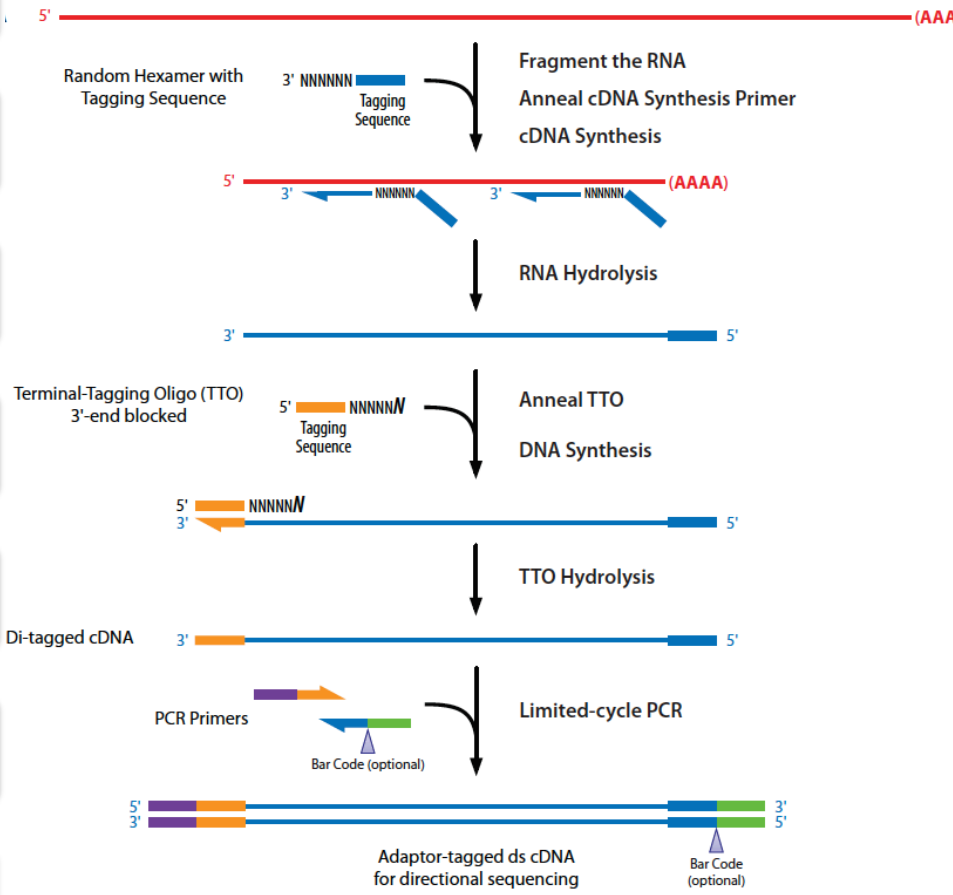
	Standard	Strand Specific
キット	• TruSeq RNA Sample Prep Kit	• Epicentre ScriptSeq™ Sample Prep Kit
スタート材料	• Total RNA • 1ug	• Poly A 、あるいは rRNA 除去処理した RNA • 50-250 ng
ワークフロー	• Poly A およびランダムプライマーを使い、2本鎖cDNA を合成	• ランダムプライマーとタグ配列を使い、2本鎖 cDNA を合成 • タグ配列でストランドを認識
利点	• 標準的な遺伝子発現解析手法 • サンプルあたりのコストが安価	• 遺伝子発現に加えてストランド情報を取得 • バクテリア、FFPEにも応用可能

Strand Specific mRNA-Seq のワークフロー

Epicentre ScriptSeq RNA Sample Prep Kit

3時間

- Poly A RNA あるいは rRNA 除去処理した RNA
- RNA の断片化
- 5’ 側にタグ配列を含むランダムプライマーを用いて 1本鎖 cDNA 合成
- 3’ 末端にタグ配列を含むターミナルオリゴを付加
- 両端にタグ配列をもつ 1本鎖 cDNA 合成
- アダプター付のタグ配列を使い低サイクルの PCR（10-12サイクル）
- 精製



illumina®

rRNA除去に使用できるキット

- Epiecntre Ribo-Zero™ rRNA Removable Kit
  - Total RNA 1～5 ug スタート
    - ヒトの場合、微量スタートキットもあり: total RNA 0.5～1 ug
  - rRNA を 99% 除去
  - 複数の生物種に対応したキット
    - Human/Mouse/Rat
    - Gram Positive, Negative Bacteria
    - Plant Leaf, Seed/Root

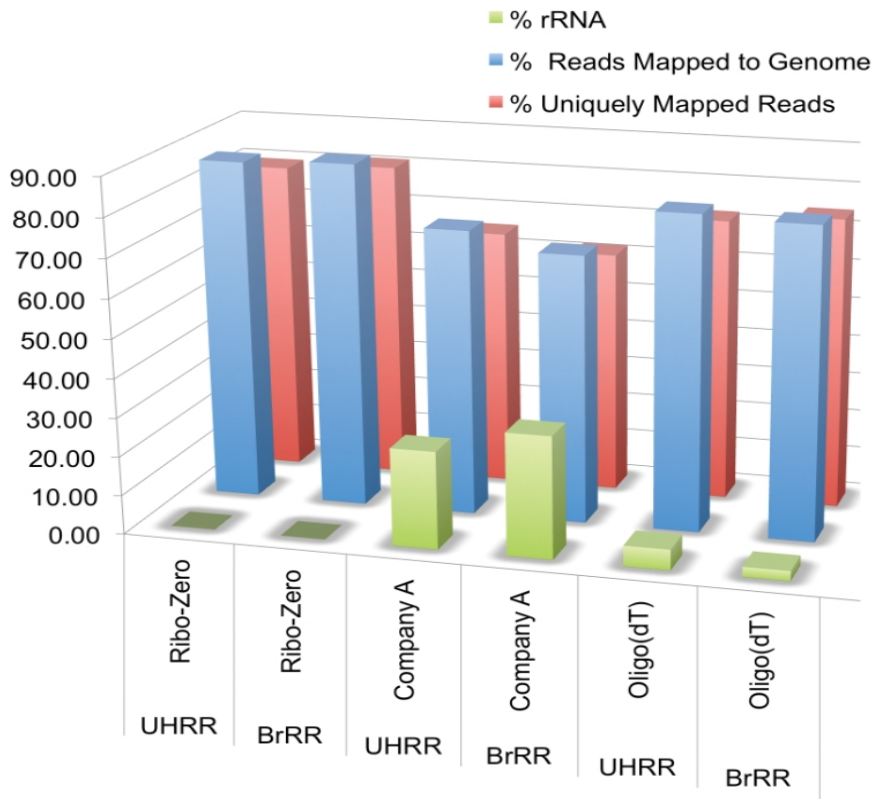
Total RNA 抽出

↓

Ribo-Zero™ rRNA Removable Kit
1～2 時間

↓

ScripSeq™ RNA Sample Prep Kit
3 時間



illumina®

サンプル調製キットのまとめ; Strand Specific mRNA-Seq

通常プロトコール

カタログ番号	製品名	キット価格	サンプルあたりの価格	問い合せ
SS10906	ScriptSeq™ mRNA-Seq Library Preparation Kit 6反応	240,000円	40,000円	エアブラウン
SS10924	ScriptSeq™ mRNA-Seq Library Preparation Kit 24反応	690,000円	28,750円	
RSBC10948	RNA-Seq Barcode Primers (Illumina社対応、12 Barcodes) 48反応	40,000円	833円	

rRNA除去用キット

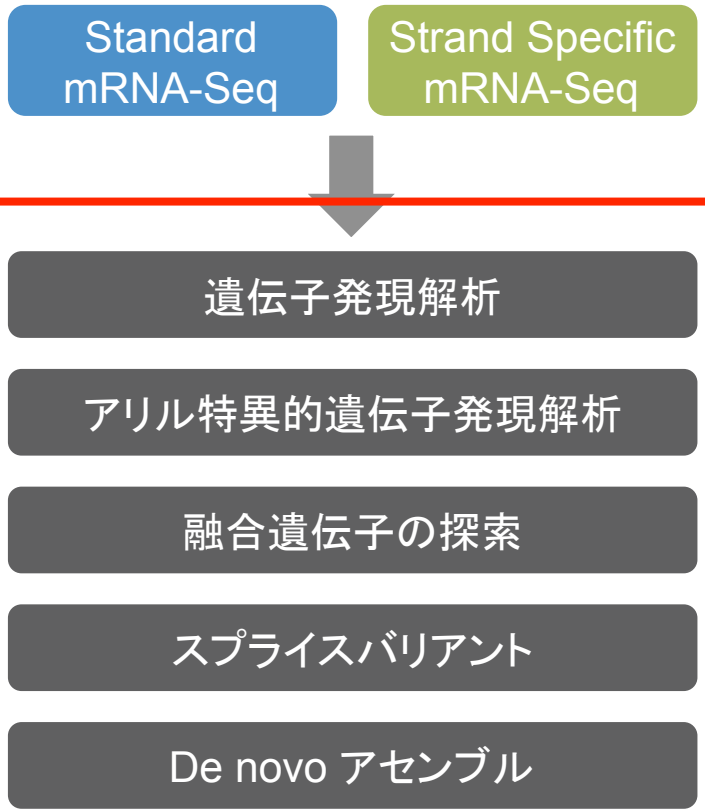
カタログ番号	製品名	キット価格	サンプルあたりの価格	問い合せ
RZH1046	Ribo-Zero™ rRNA Removal Kit (Human/Mouse/Rat) 6反応	85,000円	14,166円	エアブラウン
RZH10424	Ribo-Zero™ rRNA Removal Kit (Human/Mouse/Rat) 24反応	255,000円	10,625円	
RZH1086	Ribo-Zero™ rRNA Removal Kit (Human/Mouse/Rat) Low Input 6反応	85,000円	14,166円	
RZNB1056	Ribo-Zero™ rRNA Removal Kit (Gram-Negative Bacteria) 6反応	85,000円	14,166円	
RZNB1056	Ribo-Zero™ rRNA Removal Kit (Gram-Positive Bacteria) 6反応	85,000円	14,166円	
RZPL11016	Ribo-Zero™ rRNA Removal Kit (Plant Leaf) 6反応	85,000円	14,166円	
RZSR11036	Ribo-Zero™ rRNA Removal Kit (Plant Seed/Root) 6反応	85,000円	14,166円	

- エアブラウン株式会社　問い合せ先　03-3545-5720
- 上記製品は2011年12月末までキャンペーン対象

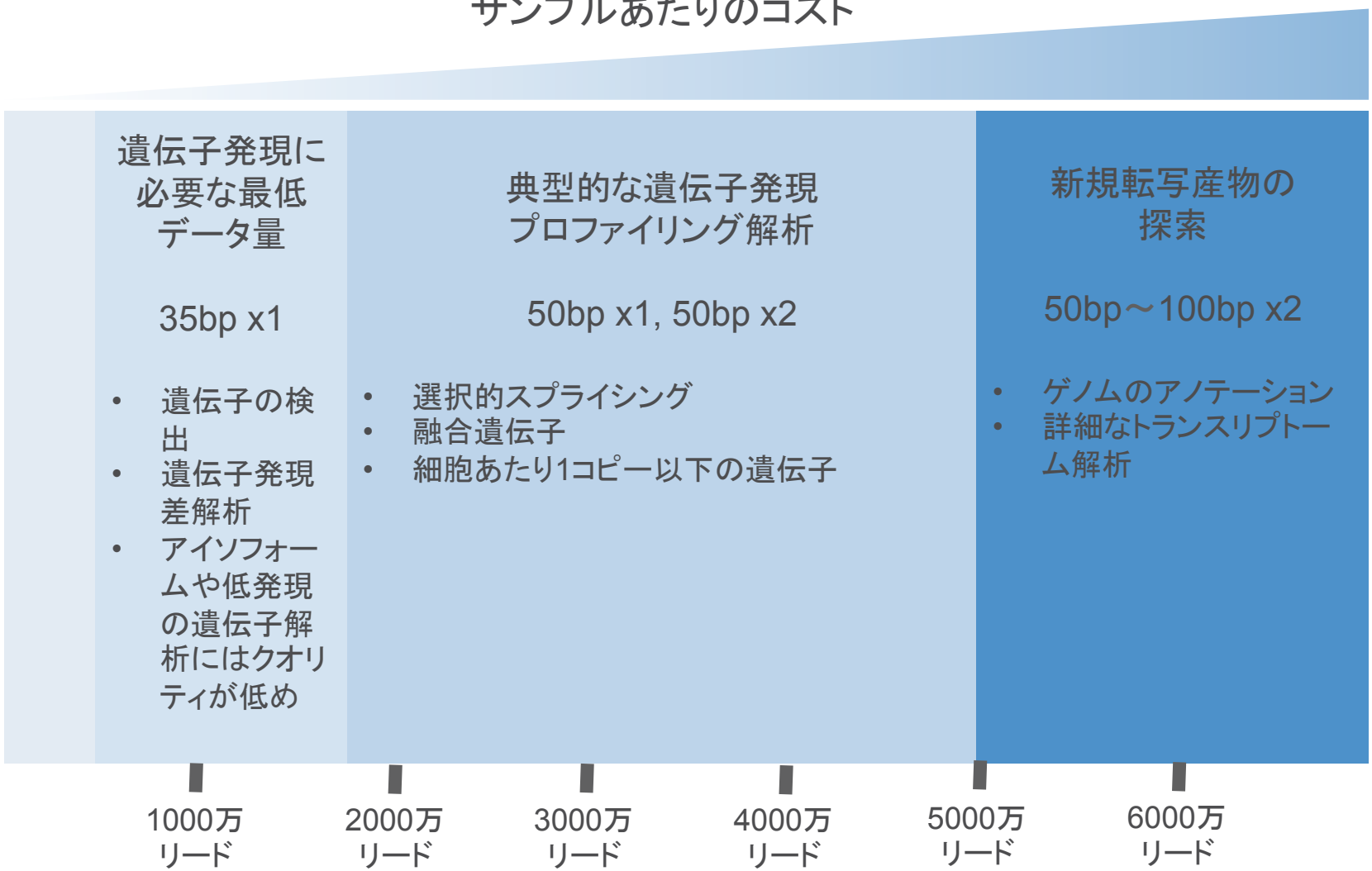
今日の内容

- サンプル調製キットのまとめ
  - Standard mRNA-Seq
  - Strand Specific mRNA-Seq

- 応用事例
  - 遺伝子発現解析
  - アリル特異的遺伝子発現
  - 融合遺伝子探索
  - スプライスバリアント
  - De novo アプリケーション



少ないリード数 & 低コスト vs. 多いリード数 & 豊富な情報



実際は次世代シーケンサーの1レーンあたりのリード数で決定している場合が多い

illumina®

HiSeq, GA 1レーンでどれぐらいのリード数が得られるか

- Genome Analyzer
  - レーンあたり 4000万 リード
  - 2サンプル／レーン　　約2000万リード／サンプル
  - 4サンプル／レーン　　約1000万リード／サンプル

- HiSeq 2000, HiSeq 1000
  - レーンあたり 1億6250万 リード
  - 2サンプル／レーン　　約8000万リード／サンプル
  - 4サンプル／レーン　　約4000万リード／サンプル
  - 8サンプル／レーン　　約2000万リード／サンプル
  - 16サンプル／レーン　　約1000万リード／サンプル

サンプルあたりのコスト試算例

- Genome Analyzer
  - レーンあたり 4000万 リード
  - 2サンプル／レーン　　約2000万リード／サンプル
  - 4サンプル／レーン　　約1000万リード／サンプル

- GA で35bp x1 を行うとすると
  - ランあたり　　約64万円
  - レーンあたり　　約8万円（8レーン／フローセル）
  - 1サンプル／レーン　　約8万円　約4000万リード／サンプル
  - 2サンプル／レーン　　約4万円　約2000万リード／サンプル
  - 4サンプル／レーン　　約2万円　約1000万リード／サンプル

- 現在のマイクロアレイ製品は約2～6万円程度
- ほぼ同じ金額でRNA-Seqができる

注） 上記はシーケンスコストのみ。追加でサンプル調製コストが必要。

アリル特異的遺伝子発現解析

マウスにおけるゲノムインプリンティング

- High Resolution Analysis of Parent-of-Origin Allelic Expression in the Mouse Brain.
  - *Science*, **329**: 643- 648, 2010

- 実験デザイン
  - Standard mRNA-Seq
  - 35bp x1
  - 約4000万リード／サンプル

- 解析のながれ
  - Novoalignでアライメント
    - ゲノムとトランスクリプトーム
  - SNPコールからアリルを特定

RESEARCH ARTICLES

High-Resolution Analysis of Parent-of-Origin Allelic Expression in the Mouse Brain

Christopher Gregg,[1,2*†] Jiangwen Zhang,[3*] Brandon Weissbourd,[1,2] Shujun Luo,[5] Gary P. Schroth,[5] David Haig,[4] Catherine Dulac[1,2†]

Genomic imprinting results in preferential expression of the paternal or maternal allele of certain genes. We have performed a genome-wide characterization of imprinting in the mouse embryonic and adult brain. This approach uncovered parent-of-origin allelic effects of more than 1300 loci. We identified parental bias in the expression of individual genes and of specific transcript isoforms, with differences between brain regions. Many imprinted genes are expressed in neural systems associated with feeding and motivated behaviors, and parental biases preferentially target genetic pathways governing metabolism and cell adhesion. We observed a preferential maternal contribution to gene expression in the developing brain and a major paternal contribution in the adult brain. Thus, parental expression bias emerges as a major mode of epigenetic regulation in the brain.

Parent-of-origin effects influence gene expression and trait inheritance in offspring. Genomic imprinting is a form of epigenetic regulation that results in the preferential expression of the paternally or maternally inherited allele of certain genes (*1*). Currently, fewer than 100 imprinted genes have been identified, and the evolutionary pressures that underlie imprinting are debated (*2*, *3*). Clinical and experimental data suggest roles for imprinting in regulating brain development and function (*4*). In humans, Prader-Willi syndrome (PWS) and Angelman syndrome (AS) result from a deletion of the paternal or maternal copy of 15q11-q13, respectively. PWS is associated with hyperphagia, stubbornness, and compulsive traits (*5*), whereas AS is associated with absent speech, happy affect, and inappropriate laughter (*6*). Further, studies of parthenogenetic (PG) and androgenetic (AG) chimeras in the mouse have suggested preferential maternal contribution to the development of the cortex, but preferential paternal contribution to the hypothalamus (*7*, *8*). Such biased roles have yet to be clearly demonstrated. Moreover, despite tantalizing reports, our understanding of the neural systems governed by imprinted genes and of the scope and features of imprinted loci expressed in the brain is very limited.

Imprinting refers to functional differences between the maternal and paternal chromosomes or alleles (*9*) and is also used more strictly to define complete allele-specific silencing (*10*). Known imprinted genes have been shown to display all-or-none and biased allelic expression according to the gene and tissue considered (*11*, *12*). We report here a genome-wide analysis of parental allelic effects involving complete silencing or parental biases in gene expression in the murine embryonic day 15 (E15) brain, and in the adult male and female cortex [medial prefrontal cortex (mPFC)] and hypothalamus [preoptic area (POA)]. Together with a companion study (*13*), our data suggest that substantial maternal and paternal biases in gene expression originate from the X chromosomes and autosomes, respectively. These results may shed light on gene regulatory processes underlying brain function, evolution, and disease.

Imprinted gene expression in the adult CNS. To gain insight into neural systems affected by imprinting, we performed an in silico study of the expression pattern of known imprinted genes in the adult brain (*14*). The expression pattern of 45 known imprinted genes was investigated across 118 distinct adult brain regions in the Allen Brain Atlas (Fig. 1 and fig. S1). A heat map based on the relative number of known imprinted genes expressed in a given brain region identified 26 out of 118 brain regions as hotspots for the expression of imprinted genes, whereas the expression hotspots of 20 randomly selected control genes with known biallelic expression were located mainly in cortical and olfactory regions and appeared entirely distinct from that of imprinted genes (Fig. 1 and fig. S1). Brain regions predicted from earlier studies to be enriched for imprinted gene expression indeed emerged as hotspots, such as the medial preoptic area (MPOA), which regulates mating, maternal behavior, and thermoregulation (*15*). From our data, aminergic systems and neural systems associated with feeding and motivated behaviors constituted the largest source of imprinting hotspots. These included the arcuate nucleus, dorsal raphe, substantia nigra pars compacta, ventral tegmental area, dorsal hypothalamic area, locus ceruleus, and nucleus accumbens (*16*, *17*). These findings enticed us to perform a more detailed and large-scale analysis to characterize and compare parent-of-origin effects governing gene expression in distinct brain regions.

A high-resolution approach to analyze imprinting. We used Illumina RNA-sequencing (RNA-Seq) technology to characterize the transcriptome of brain tissues from F_1 hybrids resulting from reciprocal crosses of CAST/EiJ (CAST) and C57BL/6J (C57) mice [F_1 initial cross (F_1i): CAST mother × C57 father; F_1 reciprocal cross (F_1r): C57 mother × CAST father]. Single-nucleotide polymorphisms (SNPs) were identified by separately sequencing the CAST and C57 transcriptomes of the original parents (or parental strains for the E15 brains), and the subsequent base calls were used to distinguish transcription from maternal and paternal alleles in F_1i and F_1r [table S1 and figs. S2 and S3 and supporting online material (SOM) (*14*)]. We characterized parent-of-origin effects governing gene expression in the E15 brain, as well as the adult male and female mPFC and POA. For the current study, male and female samples were treated as biological replicates. This approach is appropriate for the detection of parental effects that are independent of the sex of the offspring.

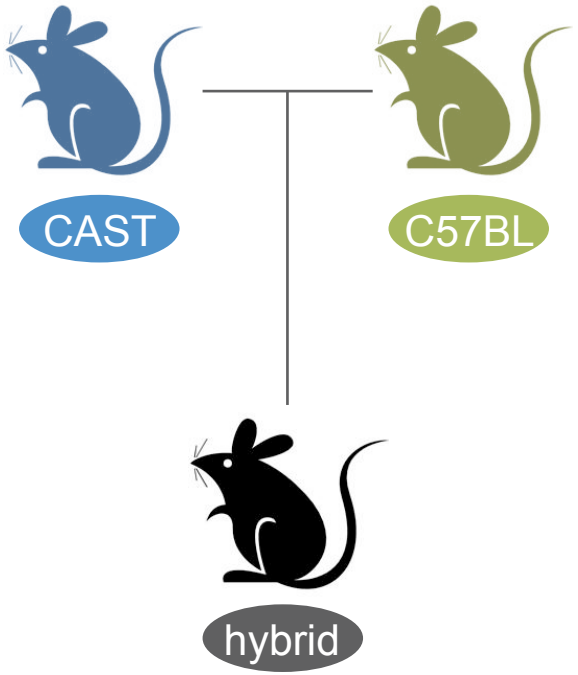
Imprinting was assessed by chi-square tests in both initial and reciprocal crosses as described in the SOM. The total number of SNP sites exhibiting a significant parent-of-origin effect was determined for a range of chi-square *P*-value cutoffs (0.001 to 0.2) and compared with the number expected by chance (Fig. 2A). We selected a cutoff of $P < 0.05$ for each cross [E15 false-discovery rate (FDR) = 0.06, POA FDR = 0.1, mPFC FDR = 0.1]. Our approach yields highly accurate and reproducible results, as demonstrated by multiple controls detailed in the SOM (*14*). Scatter plots of the –log (*P*) for the F_1i and F_1r data for each SNP site clearly indicated exclusive selection of paternally and maternally expressed loci relative to the total data set (Fig. 2B and fig. S4). Overall, SNPs identified by our approach (excluding mitochondrial and X-chromosome SNP sites) exhibited a robust parental expression bias with a mean of 87 ± 15% (mean ± SD). Parent-specific biases emerged as a continuum from the data set, which suggested that imprinting may manifest as relative allele-specific expression bias, rather than strict monoallelic transcription, or that allelic bias is cell-type specific and is partially masked by cellular heterogeneity in

[1]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA. [2]Howard Hughes Medical Institute, Harvard University, Cambridge, MA 02138, USA. [3]FAS Research Computing, Harvard University, Cambridge, MA 02138, USA. [4]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [5]Illumina, Inc., Hayward, CA 94545, USA.
*These authors contributed equally to this study.
†To whom correspondence should be addressed. E-mail: dulac@fas.harvard.edu (C.D.); cgregg@mcb.harvard.edu (C.G.)

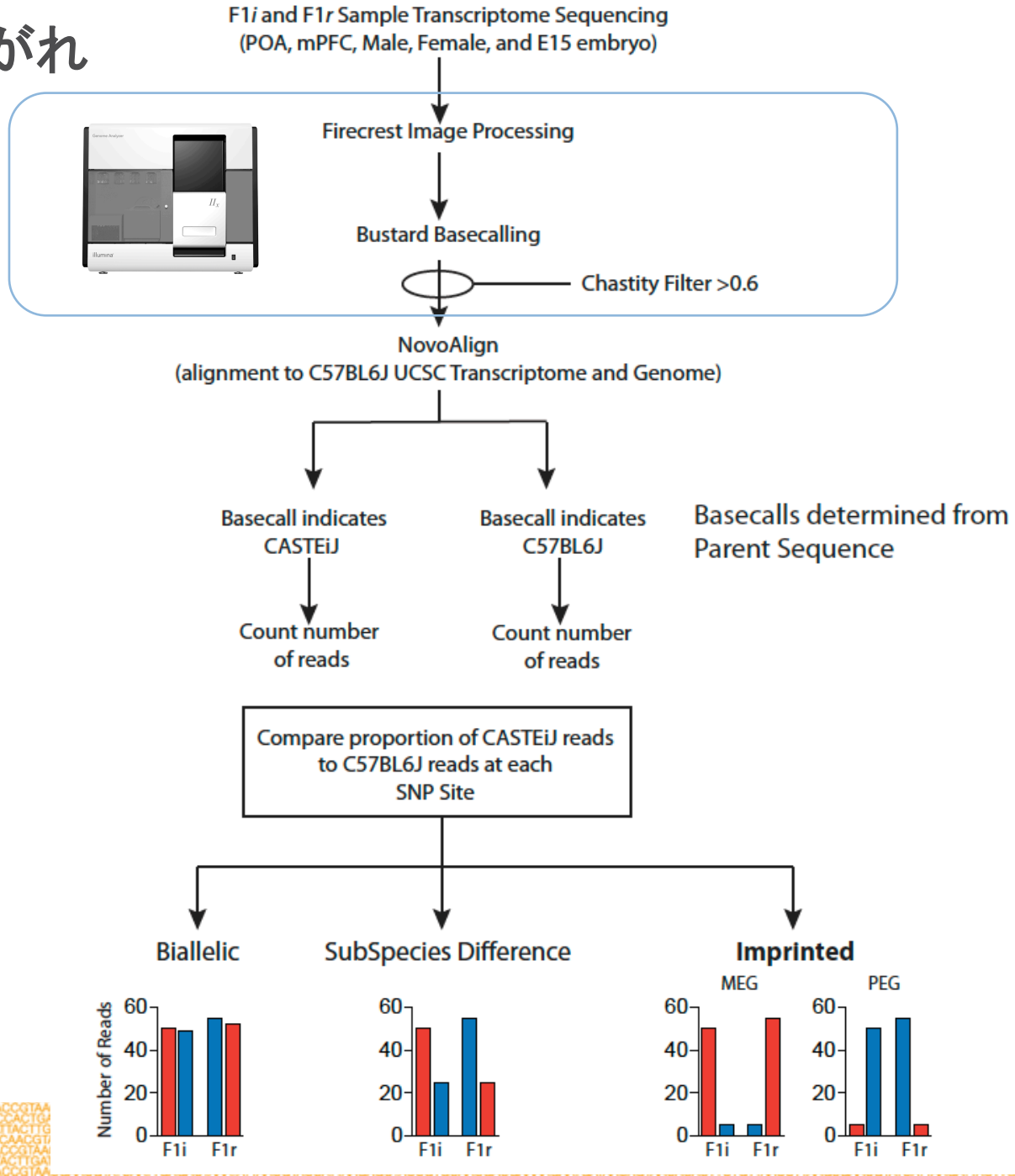
illumina®

実験デザイン

- 2種のマウスを掛け合わせ
  - CAST/EiJ と C57BL/6J を親（父、母）とするハイブリッド F1 のRNAをシーケンス
  - 各種のSNP情報から、生まれた子がどちらの親の遺伝子を発現しているかを判明



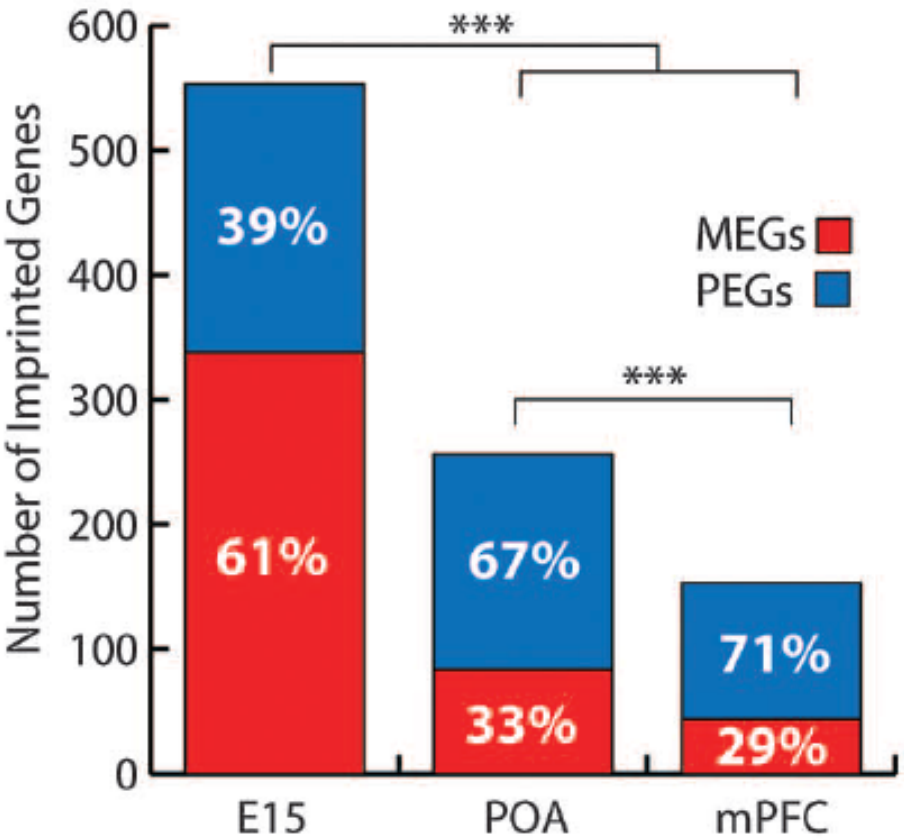
解析のながれ



illumina®

成長に応じて異なるアリル特異的発現を示す

- 発達期の脳では母親由来のアリルが遺伝子発現に貢献
- 成体脳では父親由来のアリルが遺伝子発現に貢献
- 組織ごとにアリル特異的な遺伝子発現のパターンは異なる



E15: Embryo Day 15
POA ： preoptic area 視索前野
mPFC： medial prefrontal cortex 内側前頭前皮質

同じデータセットを用いて 2報の論文を Science に投稿

High-Resolution Analysis of Parent-of-Origin Allelic Expression in the Mouse Brain

Christopher Gregg,[1,2]*† Jiangwen Zhang,[3]* Brandon Weissbourd,[1,2] Shujun Luo,[5] Gary P. Schroth,[5] David Haig,[4] Catherine Dulac[1,2]†

Genomic imprinting results in preferential expression of the paternal or maternal allele of certain genes. We have performed a genome-wide characterization of imprinting in the mouse embryonic and adult brain. This approach uncovered parent-of-origin allelic effects of more than 1300 loci. We identified parental bias in the expression of individual genes and of specific transcript isoforms, with differences between brain regions. Many imprinted genes are expressed in neural systems associated with feeding and motivated behaviors, and parental biases preferentially target genetic pathways governing metabolism and cell adhesion. We observed a preferential maternal contribution to gene expression in the developing brain and a major paternal contribution in the adult brain. Thus, parental expression bias emerges as a major mode of epigenetic regulation in the brain.

Parent-of-origin effects influence gene expression and trait inheritance in offspring. Genomic imprinting is a form of epigenetic regulation that results in the preferential expression of the paternally or maternally inherited allele of certain genes (*1*). Currently, fewer than 100 imprinted genes have been identified, and the evolutionary pressures that underlie imprinting are debated (*2*, *3*). Clinical and experimental data suggest roles for imprinting in regulating brain development and function (*4*). In humans, Prader-Willi syndrome (PWS) and Angelman syndrome (AS) result from a deletion of the paternal or maternal copy of 15q11-q13, respectively. PWS is associated with hyperphagia, stubbornness, and compulsive traits (*5*), whereas AS is associated with absent speech, happy affect, and inappropriate laughter (*6*). Further, studies of parthenogenetic (PG) and androgenetic (AG) chimeras in the mouse have suggested preferential maternal contribution to the development of the cortex, but preferential paternal contribution to the hypothalamus (*7*, *8*). Such biased roles have yet to be clearly demonstrated. Moreover, despite tantalizing reports, our understanding of the neural systems governed by imprinted genes and of the scope and features of imprinted loci expressed in the brain is very limited.

Imprinting refers to functional differences between the maternal and paternal chromosomes or alleles (*9*) and is also used more strictly to define complete allele-specific silencing (*10*). Known imprinted genes have been shown to display all-or-none and biased allelic expression according to the gene and tissue considered (*11*, *12*). We report here a genome-wide analysis of parental allelic effects involving complete silencing or parental biases in gene expression in the murine embryonic day 15 (E15) brain, and in the adult male and female cortex [medial prefrontal cortex (mPFC)] and hypothalamus [preoptic area (POA)]. Together with a companion study (*13*), our data suggest that substantial maternal and paternal biases in gene expression originate from the X chromosomes and autosomes, respectively. These results may shed light on gene regulatory processes underlying brain function, evolution, and disease.

Imprinted gene expression in the adult CNS. To gain insight into neural systems affected by imprinting, we performed an in silico study of the expression pattern of known imprinted genes in the adult brain (*14*). The expression pattern of 45 known imprinted genes was investigated across 118 distinct adult brain regions in the Allen Brain Atlas (Fig. 1 and fig. S1). A heat map based on the relative number of known imprinted genes expressed in a given brain region identified 26 out of 118 brain regions as hotspots for the expression of imprinted genes, whereas the expression hotspots of 20 randomly selected control genes with known biallelic expression were located mainly in cortical and olfactory regions and appeared entirely distinct from that of imprinted genes (Fig. 1 and fig. S1). Brain regions predicted from earlier studies to be enriched for imprinted gene expression indeed emerged as hotspots, such as the medial preoptic area (MPOA), which regulates mating, maternal behavior, and thermoregulation (*15*). From our data, aminergic systems and neural systems associated with feeding and motivated behaviors constituted the largest source of imprinting hotspots. These included the arcuate nucleus, dorsal raphe, substantia nigra pars compacta, ventral tegmental area, dorsal hypothalamic area, locus ceruleus, and nucleus accumbens (*16*, *17*). These findings enticed us to perform a more detailed and large-scale analysis to characterize and compare parent-of-origin effects governing gene expression in distinct brain regions.

A high-resolution approach to analyze imprinting. We used Illumina RNA-sequencing (RNA-Seq) technology to characterize the transcriptome of brain tissues from F_1 hybrids resulting from reciprocal crosses of CAST/EiJ (CAST) and C57BL/6J (C57) mice [F_1 initial cross (F_1i): CAST mother × C57 father; F_1 reciprocal cross (F_1r): C57 mother × CAST father]. Single-nucleotide polymorphisms (SNPs) were identified by separately sequencing the CAST and C57 transcriptomes of the original parents (or parental strains for the E15 brains), and the subsequent base calls were used to distinguish transcription from maternal and paternal alleles in F_1i and F_1r [table S1 and figs. S2 and S3 and supporting online material (SOM) (*14*)]. We characterized parent-of-origin effects governing gene expression in the E15 brain, as well as the adult male and female mPFC and POA. For the current study, male and female samples were treated as biological replicates. This approach is appropriate for the detection of parental effects that are independent of the sex of the offspring.

Imprinting was assessed by chi-square tests in both initial and reciprocal crosses as described in the SOM. The total number of SNP sites exhibiting a significant parent-of-origin effect was determined for a range of chi-square P-value cutoffs (0.001 to 0.2) and compared with the number expected by chance (Fig. 2A). We selected a cutoff of $P < 0.05$ for each cross [E15 false-discovery rate (FDR) = 0.06, POA FDR = 0.1, mPFC FDR = 0.1]. Our approach yields highly accurate and reproducible results, as demonstrated by multiple controls detailed in the SOM (*14*). Scatter plots of the –log (P) for the F_1i and F_1r data for each SNP site clearly indicated exclusive selection of paternally and maternally expressed loci relative to the total data set (Fig. 2B and fig. S4). Overall, SNPs identified by our approach (excluding mitochondrial and X-chromosome SNP sites) exhibited a robust parental expression bias with a mean of 87 ± 15% (mean ± SD). Parent-specific biases emerged as a continuum from the data set, which suggested that imprinting may manifest as relative allele-specific expression bias, rather than strict monoallelic transcription, or that allelic bias is cell-type specific and is partially masked by cellular heterogeneity in

[1]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA. [2]Howard Hughes Medical Institute, Harvard University, Cambridge, MA 02138, USA. [3]FAS Research Computing, Harvard University, Cambridge, MA 02138, USA. [4]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [5]Illumina, Inc., Hayward, CA 94545, USA.

*These authors contributed equally to this study.
†To whom correspondence should be addressed. E-mail: dulac@fas.harvard.edu (C.D.); cgregg@mcb.harvard.edu (C.G.)

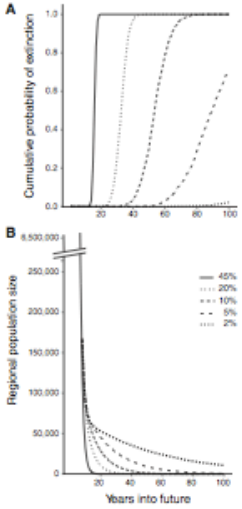


Fig. 4. (A) Cumulative probability of regional extinction of little brown myotis for five scenarios of time-dependent amelioration of disease mortality from WNS, based on matrix model simulation results. Each scenario represents predicted time-dependent declines for a specified number of years after infection and then holds the decline rate constant at either 45, 20, 10, 5, or 2% to demonstrate the impact of amelioration on the probability of extinction over the next 100 years. **(B)** Population size in each year averaged across 1000 simulations for each of the five scenarios of time-dependent amelioration of mortality from WNS.

References and Notes

1. P. Daszak, A. A. Cunningham, A. D. Hyatt, *Science* **287**, 443 (2000).
2. H. McCallum, A. Dobson, *Trends Ecol. Evol.* **10**, 190 (1995).
3. H. McCallum, *Trends Ecol. Evol.* **23**, 631 (2008).
4. A. M. Kilpatrick, C. J. Briggs, P. Daszak, *Trends Ecol. Evol.* **25**, 109 (2010).
5. L. Berger *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 9031 (1998).
6. K. R. Lips *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 3165 (2006).
7. I. Fenner, *FEMS Microbiol. Rev.* **24**, 123 (2000).
8. S. L. LaDeau, A. M. Kilpatrick, P. P. Marra, *Nature* **447**, 710 (2007).
9. D. S. Blehert *et al.*, *Science* **323**, 227 (2009).
10. A. Gargas, M. T. Trest, M. Christensen, T. J. Volk, D. S. Blehert, *Mycotaxon* **108**, 147 (2009).
11. S. J. Puechmaille *et al.*, *Emerg. Infect. Dis.* **16**, 290 (2010).
12. T. J. O'Shea, M. A. Bogan, *Monitoring Trends in Bat Populations of the United States and Territories: Problems and Prospects* (Biological Resources Discipline, Information and Technology Report USGS/BRD/ITR-2003-003, U.S. Geological Survey, Washington, DC, 2003).
13. R. Barbour, W. Davis, *Bats of America* (Univ. Press of Kentucky, Lexington, KY, USA, 1969).
14. T. H. Kunz, L. F. Lumsden, in *Bat Ecology*, T. H. Kunz, M. B. Fenton, Eds. (Univ. of Chicago Press, Chicago, IL, 2003), pp. 3–89.
15. T. H. Kunz, D. S. Reynolds, in (*12*), pp. 9–20.
16. W. H. Davis, H. B. Hitchcock, *J. Mammal.* **46**, 296 (1965).
17. D. W. Thomas, M. B. Fenton, R. M. R. Barclay, *Behav. Ecol. Sociobiol.* **6**, 129 (1979).
18. Information on materials and methods is available on *Science* Online.
19. W. F. Frick, D. S. Reynolds, T. H. Kunz, *J. Anim. Ecol.* **79**, 128 (2010).
20. E. T. Poskusny, C. Butchkoski, in *Proceedings of Bat Conservation and Mining: A Technical Interactive Forum* (Bat Conservation International and U.S. Department of the Interior, Office of Surface Mining, St. Louis, MO, 2000), pp. 159–166.
21. M. D. Tuttle, D. Hensley, *Bats* **11**, 3 (1993).
22. K. N. Geluso, J. S. Altenbach, D. E. Wilson, *Science* **194**, 184 (1976).
23. W. F. Morris, D. F. Doak, *Quantitative Conservation Biology: Theory and Practice of Population Viability Analysis* (Sinauer, Sunderland, MA, 2002).
24. F. Courchamp, T. Clutton-Brock, B. Grenfell, *Trends Ecol. Evol.* **14**, 405 (1999).
25. P. A. Stephens, W. J. Sutherland, *Trends Ecol. Evol.* **14**, 401 (1999).
26. F. Courchamp, B. Grenfell, T. Clutton-Brock, *Proc. Biol. Sci.* **266**, 557 (1999).
27. K. J. Gaston, *Science* **327**, 154 (2010).
28. G. W. Luck, G. C. Daily, P. R. Ehrlich, *Trends Ecol. Evol.* **18**, 331 (2003).
29. Funding was provided by grants from the U.S. Fish and Wildlife Service (USFWS) to W.F.F., J.F.P., D.S.R., T.H.K., and G.G.T. We thank three anonymous reviewers, J. P. Hayes, and D. F. Doak for helpful reviews and A. M. Kilpatrick for fruitful discussion. Funding for winter counts of bats at hibernacula was provided by USFWS Section 6 and State Wildlife Grants issued to the Pennsylvania Game Commission, and by Federal Aid in Wildlife Restoration Grant WE-173-G issued to the New York State Department of Environmental Conservation. Count data from hibernating colonies were kindly provided by the Connecticut Department of Environmental Protection; the Pennsylvania Game Commission; the New York Department of Environmental Conservation; Vermont Fish and Game; the Massachusetts Division of Fisheries and Wildlife; and K. Berner, State University of New York at Cobleskill. We are grateful to the many individuals who were involved in conducting annual counts of bats at hibernacula over the past 30 years. Data are available upon request from the authors.

Supporting Online Material

www.sciencemag.org/cgi/content/full/329/5992/679/DC1
Materials and Methods
Figs. S1 and S2
Tables S1 to S3
References

22 February 2010; accepted 24 May 2010
10.1126/science.1188495

Sex-Specific Parent-of-Origin Allelic Expression in the Mouse Brain

Christopher Gregg,[1,2] Jiangwen Zhang,[3] James E. Butler,[1,2] David Haig,[4] Catherine Dulac[1,2]*

Genomic imprinting results in preferential gene expression from paternally versus maternally inherited chromosomes. We used a genome-wide approach to uncover sex-specific parent-of-origin allelic effects in the adult mouse brain. Our study identified preferential selection of the maternally inherited X chromosome in glutamatergic neurons of the female cortex. Moreover, analysis of the cortex and hypothalamus identified 347 autosomal genes with sex-specific imprinting features. In the hypothalamus, sex-specific imprinted genes were mostly found in females, which suggests parental influence over the hypothalamic function of daughters. We show that *interleukin-18*, a gene linked to diseases with sex-specific prevalence, is subject to complex, regional, and sex-specific parental effects in the brain. Parent-of-origin effects thus provide new avenues for investigation of sexual dimorphism in brain function and disease.

Genomic imprinting is an epigenetic mode of gene regulation involving preferential expression of the paternally or maternally inherited allele (*1*). Sexual dimorphism is a central characteristic of mammalian brain function and behavior that influences major neurological diseases in humans (*2*). Here we address the potential existence of differential genomic imprinting in the brain according to the sex of individuals. Imprinting refers to gene expression differences between maternal and paternal chromosomes (*3*) and is also used more strictly to define complete allele-specific silencing (*4*). Our analysis encompasses sex differences in parent-

[1]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA. [2]Howard Hughes Medical Institute, Harvard University, Cambridge, MA 02138, USA. [3]FAS Research Computing, Harvard University, Cambridge, MA 02138, USA. [4]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA.
*To whom correspondence should be addressed. E-mail: dulac@fas.harvard.edu

illumina®

アリル特異的遺伝子発現の文献

- Allele-specific expression assays using Solexa. *BMC Genomics* 2009, **10:**422
- Identification of transcriptome SNPs between Xiphophorus lines and species for assessing allele specific gene expression within F(1) interspecies hybrids. *Comp Biochem Physiol C Toxicol Pharmacol. 2011 Apr 3.*
- Genome-wide identification of allele-specific expression (ASE) in response to Mareks disease virus infection using next generation sequencing. *BMC Proc. 2011; 5(Suppl 4): S14.*

融合遺伝子の探索

融合遺伝子探索にはシングル vs. ペアどちらが有利？

- Chimeric transcript discovery by paired-end transcriptome sequencing
  - Proc Natl Acad Sci U S A. 2009 Jul 28;106(30):12353-8.
- 実験デザイン
  - Standard mRNA-Seq
  - 50bp x2 と 100bp x1 の比較
  - 700-5300万リード
- 解析のながれ
  - hg18, RefSeqにマッピング（ELAND）
  - 転写産物、ミトコンドリア、rRNA、コントロールにマップしたリードは排除
  - キメラ候補、マップできないリードを解析

PNAS

Chimeric transcript discovery by paired-end transcriptome sequencing

Christopher A. Maher[a,b], Nallasivam Palanisamy[a,b], John C. Brenner[a,b], Xuhong Cao[a,c], Shanker Kalyana-Sundaram[a,b], Shujun Luo[d], Irina Khrebtukova[d], Terrence R. Barrette[a,b], Catherine Grasso[a,b], Jindan Yu[a,b], Robert J. Lonigro[a,b], Gary Schroth[d], Chandan Kumar-Sinha[a,b], and Arul M. Chinnaiyan[a,b,c,e,f,1]

[a]Michigan Center for Translational Pathology, Ann Arbor, MI 48109; Departments of [b]Pathology and [c]Urology, University of Michigan, Ann Arbor, MI 48109; [e]Howard Hughes Medical Institute and [f]Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, MI 48109; and [d]Illumina Inc., 25861 Industrial Boulevard, Hayward, CA 94545

Communicated by David Ginsburg, University of Michigan Medical School, Ann Arbor, MI, May 4, 2009 (received for review March 16, 2009)

Recurrent gene fusions are a prevalent class of mutations arising from the juxtaposition of 2 distinct regions, which can generate novel functional transcripts that could serve as valuable therapeutic targets in cancer. Therefore, we aim to establish a sensitive, high-throughput methodology to comprehensively catalog functional gene fusions in cancer by evaluating a paired-end transcriptome sequencing strategy. Not only did a paired-end approach provide a greater dynamic range in comparison with single read based approaches, but it clearly distinguished the high-level "driving" gene fusions, such as *BCR-ABL1* and *TMPRSS2-ERG*, from potential lower level "passenger" gene fusions. Also, the comprehensiveness of a paired-end approach enabled the discovery of 12 previously undescribed gene fusions in 4 commonly used cell lines that eluded previous approaches. Using the paired-end transcriptome sequencing approach, we observed read-through mRNA chimeras, tissue-type restricted chimeras, converging transcripts, diverging transcripts, and overlapping mRNA transcripts. Last, we successfully used paired-end transcriptome sequencing to detect previously undescribed ETS gene fusions in prostate tumors. Together, this study establishes a highly specific and sensitive approach for accurately and comprehensively cataloguing chimeras within a sample using paired-end transcriptome sequencing.

bioinformatics | gene fusions | prostate cancer | breast cancer | RNA-Seq

One of the most common classes of genetic alterations is gene fusions, resulting from chromosomal rearrangements (1). Intriguingly, >80% of all known gene fusions are attributed to leukemias, lymphomas, and bone and soft tissue sarcomas that account for only 10% of all human cancers. In contrast, common epithelial cancers, which account for 80% of cancer-related deaths, can only be attributed to 10% of known recurrent gene fusions (2–4). However, the recent discovery of a recurrent gene fusion, *TMPRSS2-ERG*, in a majority of prostate cancers (5, 6), and *EML4-ALK* in non-small-cell lung cancer (NSCLC) (7), has expanded the realm of gene fusions as an oncogenic mechanism in common solid cancers. Also, the restricted expression of gene fusions to cancer cells makes them desirable therapeutic targets. One successful example is imatinib mesylate, or Gleevec, that targets *BCR-ABL1* in chronic myeloid leukemia (CML) (8–10). Therefore, the identification of novel gene fusions in a broad range of cancers is of enormous therapeutic significance.

The lack of known gene fusions in epithelial cancers has been attributed to their clonal heterogeneity and to the technical limitations of cytogenetic analysis, spectral karyotyping, FISH, and microarray-based comparative genomic hybridization (aCGH). Not surprisingly, *TMPRSS2-ERG* was discovered by circumventing these limitations through bioinformatics analysis of gene expression data to nominate genes with marked overexpression, or outliers, a signature of a fusion event (6). Building on this success, more recent strategies have adopted unbiased high-throughput approaches, with increased resolution, for genome-wide detection of chromosomal rearrangements in cancer involving BAC end sequencing (11), fosmid paired-end sequences (12), serial analysis of gene expression (SAGE)-like sequencing (13), and next-generation DNA sequencing (14). Despite unveiling many novel genomic rearrangements, solid tumors accumulate multiple nonspecific aberrations throughout tumor progression; thus, making causal and driver aberrations indistinguishable from secondary and insignificant mutations, respectively.

The deep unbiased view of a cancer cell enabled by massively parallel transcriptome sequencing has greatly facilitated gene fusion discovery. As shown in our previous work, integrating long and short read transcriptome sequencing technologies was an effective approach for enriching "expressed" fusion transcripts (15). However, despite the success of this methodology, it required substantial overhead to leverage 2 sequencing platforms. Therefore, in this study, we adopted a single platform paired-end strategy to comprehensively elucidate novel chimeric events in cancer transcriptomes. Not only was using this single platform more economical, but it allowed us to more comprehensively map chimeric mRNA, hone in on driver gene fusion products due to its quantitative nature, and observe rare classes of transcripts that were overlapping, diverging, or converging.

Results

Chimera Discovery via Paired-End Transcriptome Sequencing. Here, we employ transcriptome sequencing to restrict chimera nominations to "expressed sequences," thus, enriching for potentially functional mutations. To evaluate massively parallel paired-end transcriptome sequencing to identify novel gene fusions, we generated cDNA libraries from the prostate cancer cell line VCaP, CML cell line K562, universal human reference total RNA (UHR; Stratagene), and human brain reference (HBR) total RNA (Ambion). Using the Illumina Genome Analyzer II, we generated 16.9 million VCaP, 20.7 million K562, 25.5 million UHR, and 23.6 million HBR transcriptome mate pairs (2 × 50 nt). The mate pairs were mapped against the transcriptome and categorized as (*i*) mapping to same gene, (*ii*) mapping to different genes (chimera candidates), (*iii*) nonmapping, (*iv*) mitochondrial, (*v*) quality control, or (*vi*) ribosomal (Table S1). Overall, the chimera candidates represent a minor fraction of the mate pairs, comprising ≈<1% of the reads for each sample.

We believe that a paired-end strategy offers multiple advantages over single read based approaches such as alleviating the reliance on sequencing the reads traversing the fusion junction, increased coverage provided by sequencing reads from the ends of a tran-

Author contributions: C.A.M. and A.M.C. designed research; C.A.M., N.P., J.C.B., X.C., S.L., I.K., T.R.B., R.J.L., G.S., C.K.-S., and A.M.C. performed research; C.A.M., S.L., I.K., R.J.L., and G.S. contributed new reagents/analytic tools; C.A.M., N.P., J.C.B., S.K.-S., C.G., J.Y., R.J.L., G.S., C.K.-S., and A.M.C. analyzed data; and C.A.M., N.P., X.C., C.K.-S., and A.M.C. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: arul@umich.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0904720106/DCSupplemental.

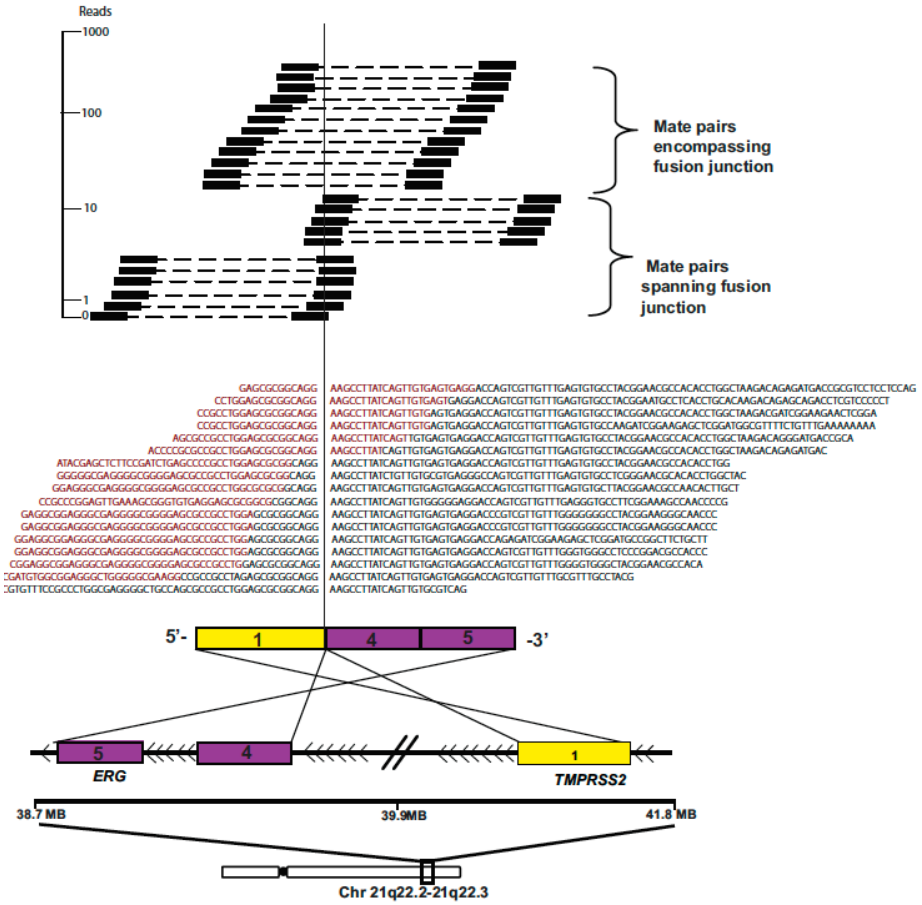
CELL BIOLOGY

illumina®

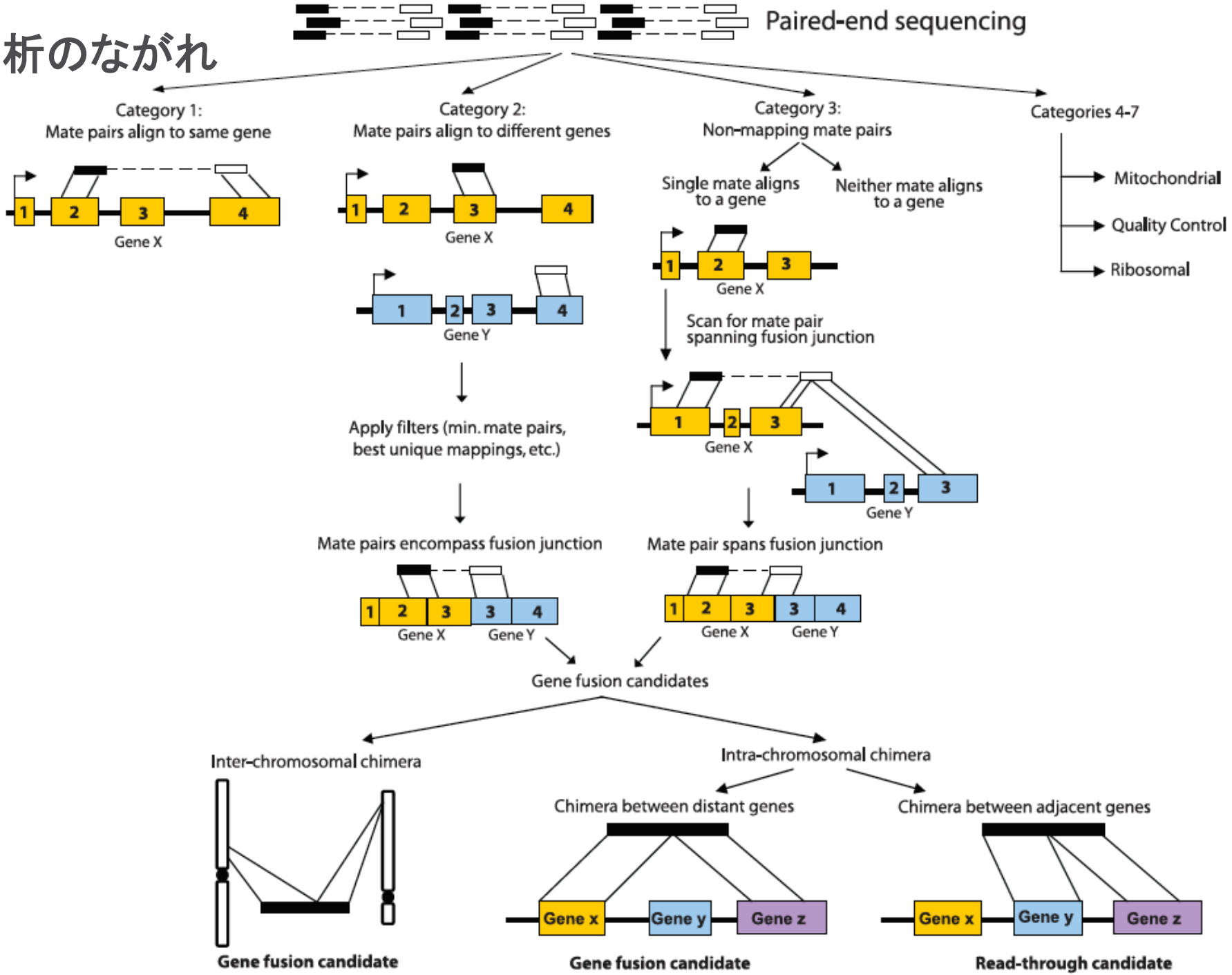
実験デザイン

- 4種類のセルライン（前立腺癌、CML、UHRR、Brain）にて 50bp x2, 100bp x1 を行い融合遺伝子の検出
- リード数

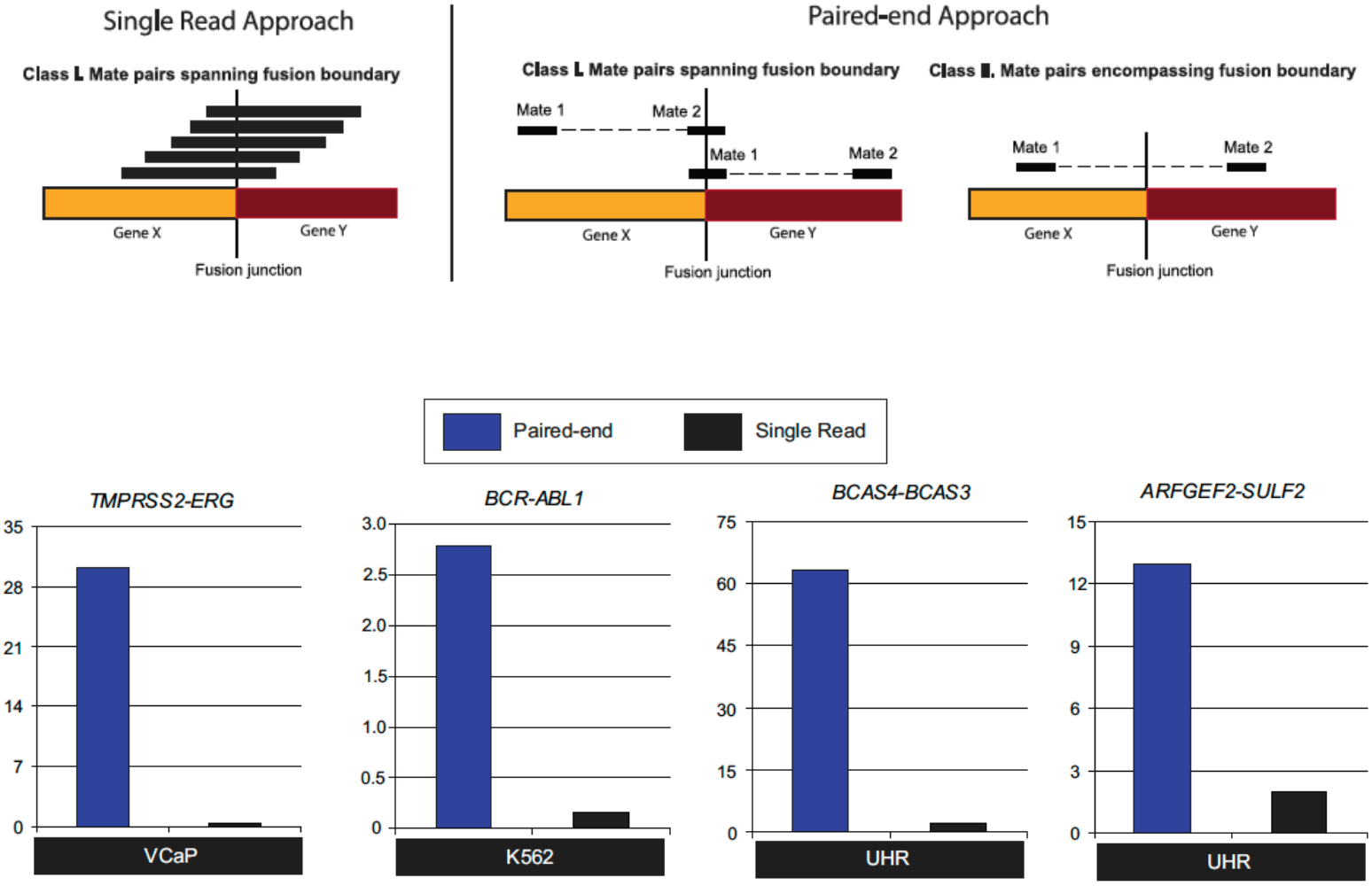
	50bp x2	100bp x1
VCaP	1690万	700万
K562	2070万	
Human Ref	2250万	5940万
Brain	2360万	5300万



解析のながれ


Paired-end sequencing
Category 1:
Mate pairs align to same gene
Gene X
Category 2:
Mate pairs align to different genes
Gene X
Gene Y
Apply filters (min. mate pairs, best unique mappings, etc.)
Mate pairs encompass fusion junction
Gene X
Gene Y
Category 3:
Non-mapping mate pairs
Single mate aligns to a gene
Neither mate aligns to a gene
Gene X
Scan for mate pair spanning fusion junction
Gene X
Gene Y
Mate pair spans fusion junction
Gene X
Gene Y
Categories 4-7
Mitochondrial
Quality Control
Ribosomal
Gene fusion candidates
Inter-chromosomal chimera
Gene fusion candidate
Intra-chromosomal chimera
Chimera between distant genes
Gene x
Gene y
Gene z
Gene fusion candidate
Chimera between adjacent genes
Gene x
Gene y
Gene z
Read-through candidate

ペアエンドの方が既知融合遺伝子を高い感度で検出



illumina®

融合遺伝子の文献

Cancer

- MHC class II transactivator CIITA is a recurrent gene fusion partner in lymphoid cancers. *Nature. 2011 Mar 17;471(7338):377-81.*
- Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC Med Genomics. 2011 Jan 24;4:11.*
- N-myc downstream regulated gene 1 (NDRG1) is fused to ERG in prostate cancer. *Neoplasia. 2009 Aug;11(8):804-11.*
- Transcriptome sequencing to detect gene fusions in cancer. *Nature. 2009 Mar 5;458 (7234):97-101.*
- Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. *JAMA. 2011 Apr 20;305(15):1577-84.*

Data Analysis

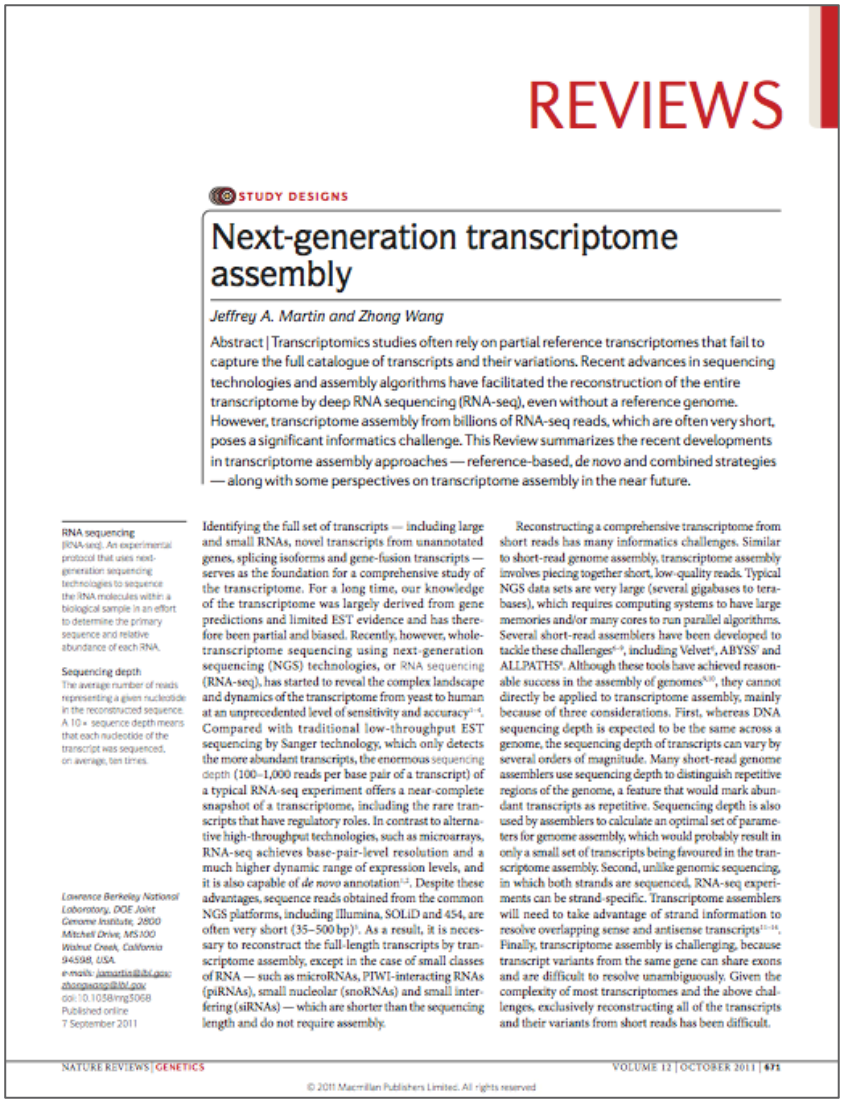
- Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. *Bioinformatics. 2011 Apr 15;27(8):1068-75.*
- ChimerDB 2.0--a knowledgebase for fusion genes updated. *Nucleic Acids Res. 2010 Jan; 38*

トランスクリプトームのアセンブル戦略

トランスクリプトームを用いたアセンブル戦略

- 方法は3つ
  - リファレンス配列を使用（Isoform）
  - De novo
  - リファレンス配列とde novoを両用

- 参考資料：
  Next-generation transcriptome assembly
  - Nature Reviews VOLUME 12 | OCTOBER 2011 | 671



REVIEWS

STUDY DESIGNS

Next-generation transcriptome assembly

Jeffrey A. Martin and Zhong Wang

Abstract | Transcriptomics studies often rely on partial reference transcriptomes that fail to capture the full catalogue of transcripts and their variations. Recent advances in sequencing technologies and assembly algorithms have facilitated the reconstruction of the entire transcriptome by deep RNA sequencing (RNA-seq), even without a reference genome. However, transcriptome assembly from billions of RNA-seq reads, which are often very short, poses a significant informatics challenge. This Review summarizes the recent developments in transcriptome assembly approaches — reference-based, *de novo* and combined strategies — along with some perspectives on transcriptome assembly in the near future.

RNA sequencing (RNA-seq). An experimental protocol that uses next-generation sequencing technologies to sequence the RNA molecules within a biological sample in an effort to determine the primary sequence and relative abundance of each RNA.

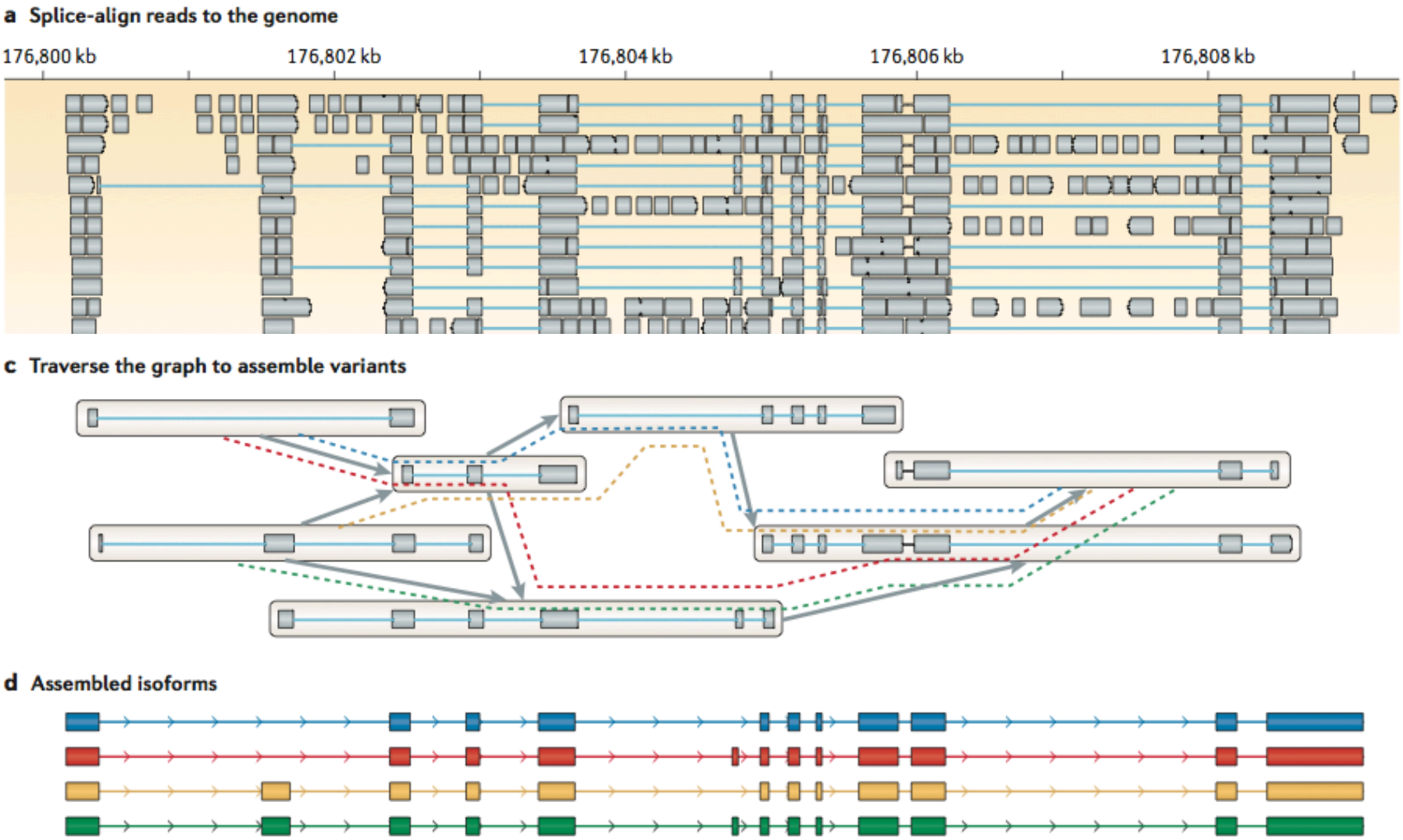
Sequencing depth. The average number of reads representing a given nucleotide in the reconstructed sequence. A 10× sequence depth means that each nucleotide of the transcript was sequenced, on average, ten times.

Lawrence Berkeley National Laboratory, DOE Joint Genome Institute, 2800 Mitchell Drive, MS 100 Walnut Creek, California 94598, USA.
e-mails: jamartin@lbl.gov; zhongwang@lbl.gov
doi:10.1038/nrg3068
Published online 7 September 2011

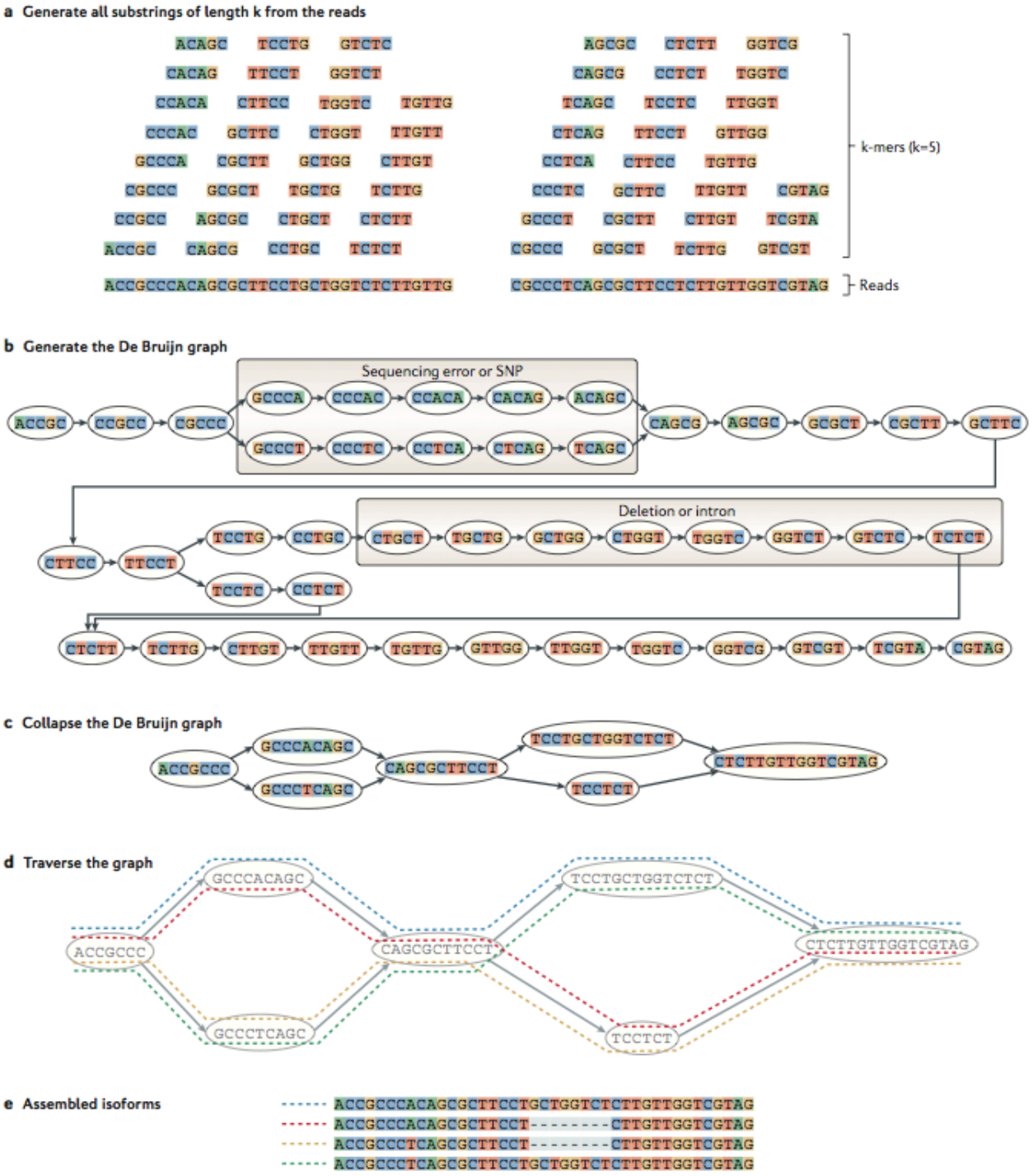
Identifying the full set of transcripts — including large and small RNAs, novel transcripts from unannotated genes, splicing isoforms and gene-fusion transcripts — serves as the foundation for a comprehensive study of the transcriptome. For a long time, our knowledge of the transcriptome was largely derived from gene predictions and limited EST evidence and has therefore been partial and biased. Recently, however, whole-transcriptome sequencing using next-generation sequencing (NGS) technologies, or RNA sequencing (RNA-seq), has started to reveal the complex landscape and dynamics of the transcriptome from yeast to human at an unprecedented level of sensitivity and accuracy[1–4]. Compared with traditional low-throughput EST sequencing by Sanger technology, which only detects the more abundant transcripts, the enormous sequencing depth (100–1,000 reads per base pair of a transcript) of a typical RNA-seq experiment offers a near-complete snapshot of a transcriptome, including the rare transcripts that have regulatory roles. In contrast to alternative high-throughput technologies, such as microarrays, RNA-seq achieves base-pair-level resolution and a much higher dynamic range of expression levels, and it is also capable of *de novo* annotation[1,2]. Despite these advantages, sequence reads obtained from the common NGS platforms, including Illumina, SOLiD and 454, are often very short (35–500 bp)[1]. As a result, it is necessary to reconstruct the full-length transcripts by transcriptome assembly, except in the case of small classes of RNA — such as microRNAs, PIWI-interacting RNAs (piRNAs), small nucleolar (snoRNAs) and small interfering (siRNAs) — which are shorter than the sequencing length and do not require assembly.

Reconstructing a comprehensive transcriptome from short reads has many informatics challenges. Similar to short-read genome assembly, transcriptome assembly involves piecing together short, low-quality reads. Typical NGS data sets are very large (several gigabases to terabases), which requires computing systems to have large memories and/or many cores to run parallel algorithms. Several short-read assemblers have been developed to tackle these challenges[5–9], including Velvet[6], ABySS[7] and ALLPATHS[8]. Although these tools have achieved reasonable success in the assembly of genomes[8,10], they cannot directly be applied to transcriptome assembly, mainly because of three considerations. First, whereas DNA sequencing depth is expected to be the same across a genome, the sequencing depth of transcripts can vary by several orders of magnitude. Many short-read genome assemblers use sequencing depth to distinguish repetitive regions of the genome, a feature that would mark abundant transcripts as repetitive. Sequencing depth is also used by assemblers to calculate an optimal set of parameters for genome assembly, which would probably result in only a small set of transcripts being favoured in the transcriptome assembly. Second, unlike genomic sequencing, in which both strands are sequenced, RNA-seq experiments can be strand-specific. Transcriptome assemblers will need to take advantage of strand information to resolve overlapping sense and antisense transcripts[11–14]. Finally, transcriptome assembly is challenging, because transcript variants from the same gene can share exons and are difficult to resolve unambiguously. Given the complexity of most transcriptomes and the above challenges, exclusively reconstructing all of the transcripts and their variants from short reads has been difficult.

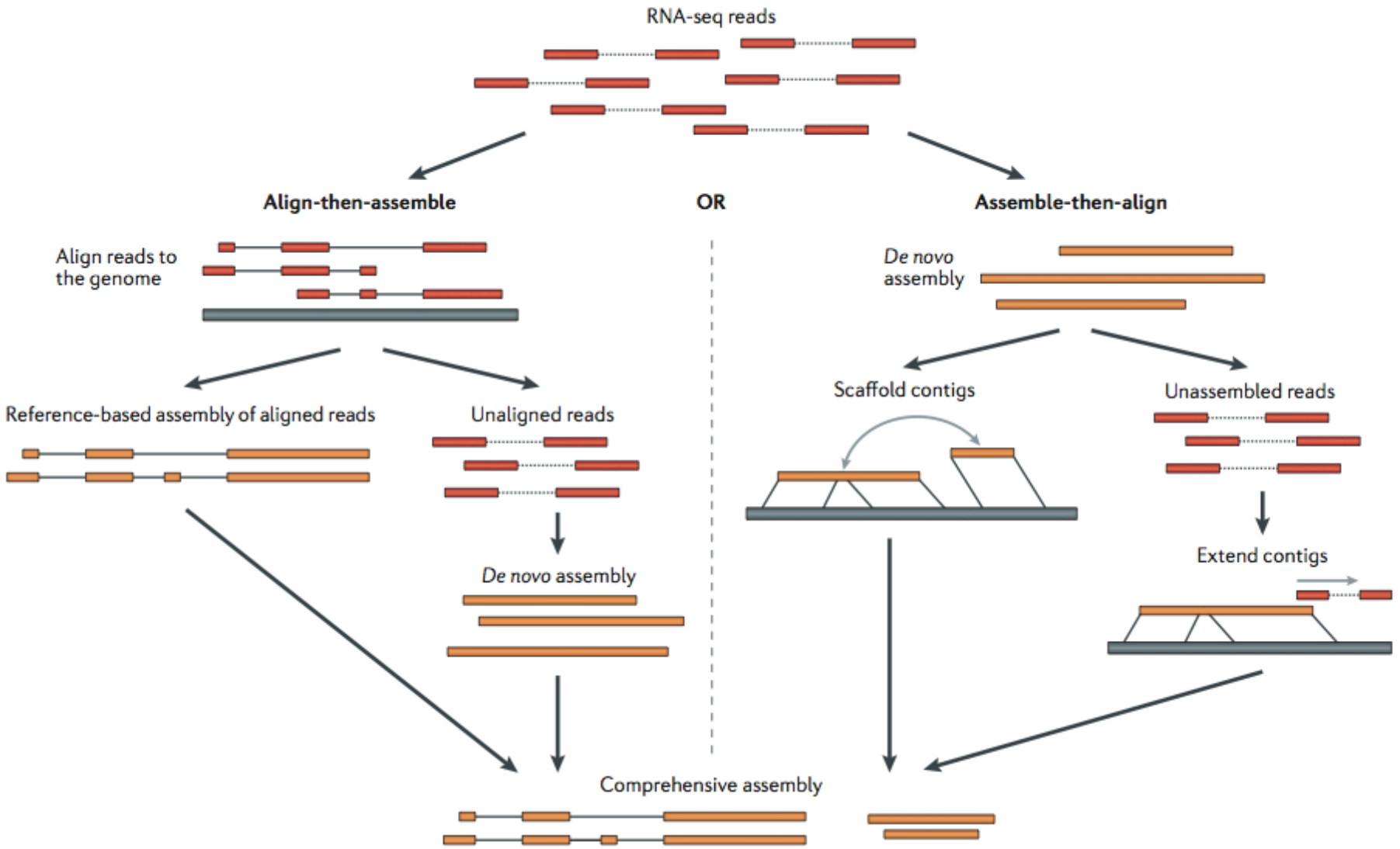
リファレンス配列を用いたアプローチ



De novo アプローチ



リファレンス配列とde novoを用いたアプローチ



トランスクリプトームのアセンブル戦略
スプライスバリアント解析

スプライスごとの発現比較

- Alternative isoform regulation in human tissue transcriptomes
  - doi:10.1038/nature07509

- 実験デザイン
  - Standard mRNA-Seq
  - 10種類の組織、5種類のセルラインから合計4億リード（1200-2900万リード／サンプル）
  - 32bp
  - データ解析のながれ
  - 予測および既知のスプライスジャンクションにヒットするリードを解析

nature Vol 456|27 November 2008|doi:10.1038/nature07509

ARTICLES

Alternative isoform regulation in human tissue transcriptomes

Eric T. Wang[1,2]*, Rickard Sandberg[1,3]*, Shujun Luo[4], Irina Khrebtukova[4], Lu Zhang[4], Christine Mayr[5], Stephen F. Kingsmore[6], Gary P. Schroth[4] & Christopher B. Burge[1]

Through alternative processing of pre-messenger RNAs, individual mammalian genes often produce multiple mRNA and protein isoforms that may have related, distinct or even opposing functions. Here we report an in-depth analysis of 15 diverse human tissue and cell line transcriptomes on the basis of deep sequencing of complementary DNA fragments, yielding a digital inventory of gene and mRNA isoform expression. Analyses in which sequence reads are mapped to exon–exon junctions indicated that 92–94% of human genes undergo alternative splicing, ~86% with a minor isoform frequency of 15% or more. Differences in isoform-specific read densities indicated that most alternative splicing and alternative cleavage and polyadenylation events vary between tissues, whereas variation between individuals was approximately twofold to threefold less common. Extreme or 'switch-like' regulation of splicing between tissues was associated with increased sequence conservation in regulatory regions and with generation of full-length open reading frames. Patterns of alternative splicing and alternative cleavage and polyadenylation were strongly correlated across tissues, suggesting coordinated regulation of these processes, and sequence conservation of a subset of known regulatory motifs in both alternative introns and 3′ untranslated regions suggested common involvement of specific factors in tissue-level regulation of both splicing and polyadenylation.

The mRNA and protein isoforms produced by alternative processing of primary RNA transcripts may differ in structure, function, localization or other properties[1,2]. Alternative splicing in particular is known to affect more than half of all human genes, and has been proposed as a primary driver of the evolution of phenotypic complexity in mammals[3,4]. However, assessment of the extent of differences in mRNA isoform expression between tissues has presented substantial technical challenges[5]. Studies using expressed sequence tags have yielded relatively low estimates of tissue specificity, but have limited statistical power to detect differences in isoform levels[6–8]. Microarray analyses have achieved more consistent coverage of tissues[9], but are constrained in their ability to distinguish closely related mRNA isoforms. High-throughput sequencing technologies have the potential to circumvent these limitations by generating high average coverage of mRNAs across tissues while using direct sequencing rather than hybridization to distinguish and quantify mRNA isoforms[10,11].

Tissue-specific alternative splicing is usually regulated by a combination of tissue-specific and ubiquitously expressed RNA-binding factors that interact with *cis*-acting RNA elements to influence spliceosome assembly at nearby splice sites[1,2]. Many factors can both activate and repress splicing in different contexts, with activity often summarizable by an 'RNA map' describing dependence on the location of binding relative to that of core spliceosomal components[12,13].

A digital inventory of mRNA isoforms

To assess gene and alternative mRNA isoform expression, the mRNA-Seq protocol (Supplementary Methods) was used to amplify and sequence between 12 million and 29 million 32-base-pair (bp) cDNA fragments from ten diverse human tissues and five mammary epithelial or breast cancer cell lines, generating over 400 million reads in total (Supplementary Fig. 1a). Tissue samples were derived from single anonymous unrelated individuals of both sexes; for one tissue, cerebellar cortex, samples from six unrelated men were analysed to assess variation between individuals (Supplementary Table 1). In total, ~60% of reads mapped uniquely to the genome, allowing up to 2 mismatches, and an additional 4% mapped uniquely to splice junctions. Thus, about two-thirds of reads could be assigned unambiguously to individual genes; the frequency of mapping to incorrect genomic locations was estimated to be ~0.1% (Supplementary Table 2).

Read density (coverage) was over 100-fold higher in exons than in introns or intergenic regions (Supplementary Fig. 1c), and only ~3% of reads mapped to ribosomal RNA genes, indicating that most reads derived from mature mRNA. Comparison of relative mRNA-Seq read densities to published quantitative polymerase chain reaction with reverse transcription (RT–PCR) measurements for 787 genes in two reference RNA samples[14] yielded a nearly linear relationship across ~5 orders of magnitude (Supplementary Fig. 1d), indicating that mRNA-Seq read counts give accurate relative gene expression measurements across a very broad dynamic range[10].

Alternative splicing is nearly universal

The mRNA-Seq data were used to assess the expression of alternative transcript isoforms in human genes, as illustrated for the mitochondrial phosphate transporter gene *SLC25A3* in Fig. 1a. Exons 3A and 3B of this gene are 'mutually exclusive exons' (MXEs), meaning that transcripts from this gene contain one or the other of these exons, but not both. Much greater read coverage of exon 3A was seen in heart and skeletal muscle, with almost exclusive coverage of exon 3B in

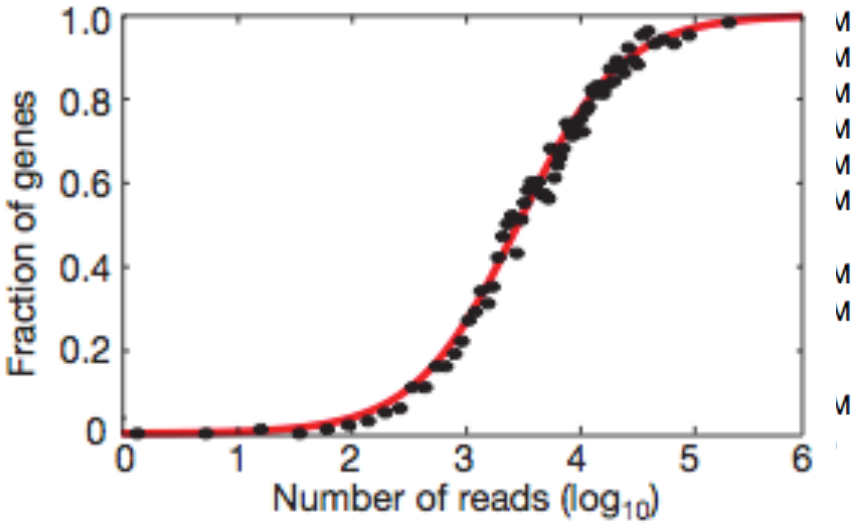
[1]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [2]Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts 02139, USA. [3]Department of Cell and Molecular Biology, Karolinska Institutet, 171 77 Stockholm, Sweden. [4]Illumina Inc., 25861 Industrial Boulevard, Hayward, California 94545, USA. [5]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA. [6]National Center for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, New Mexico 87505, USA.
*These authors contributed equally to this work.

illumina®

スプラスバリアントの検出のながれ

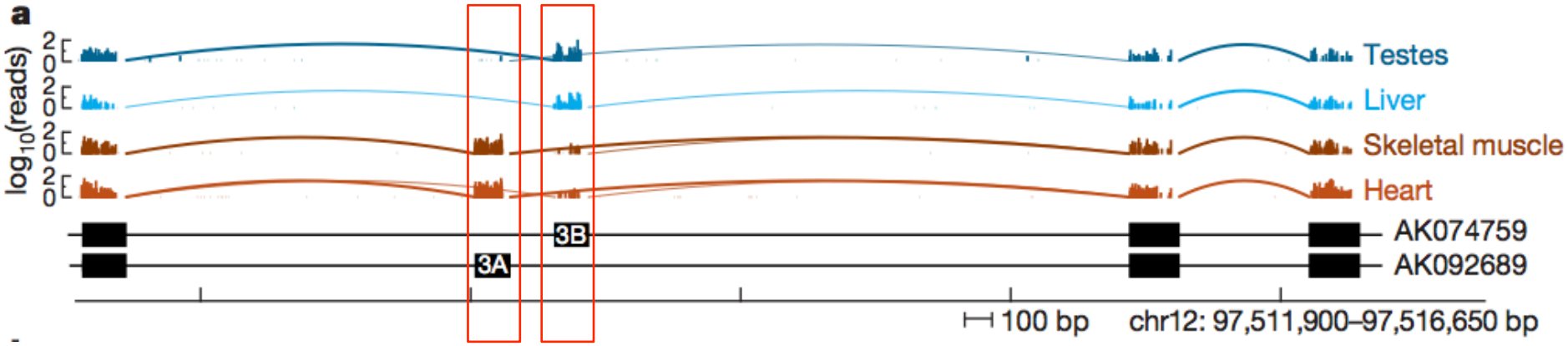
- RefSeqのうち、94%はマルチエクソンから構成される遺伝子
- スプライスを起こしている遺伝子の検出
  - 複数回およびマップ地点が異なるリードを確認
  - 5'と3'のエクソンの組み合わせが異なる
  - ジャンクションあたり2リード以上ヒットする
- 上記94%のRefSeqについて、アイソフォームをもつ割合
  - 10種類の組織では98%
  - 5種類のセルラインを追加するとほぼ100%

		Genomic Reads		Junction Reads
Samples	No. Reads	unique	non-unique	unique
Adipose	27,752,231	17.6 M	4.8 M	1.4 M
Brain	17,246,957	11.0 M	3.2 M	0.6 M
Breast	16,120,746	10.6 M	2.9 M	0.8 M
Colon	28,435,996	17.7 M	5.5 M	1.3 M
Heart	20,169,301	11.3 M	5.1 M	0.7 M
Liver	18,517,121	11.5 M	3.6 M	1.0 M
Lymph node	27,492,254	15.8 M	6.6 M	1.4 M
Skeletal muscle	22,640,454	14.4 M	4.0 M	1.3 M
Testes	27,303,938	18.6 M	4.1 M	1.6 M
BT474	18,424,533	11.5 M	3.2 M	0.8 M
HME	19,657,452	12.4 M	3.7 M	1.2 M
MB435	18,610,758	12.5 M	3.2 M	1.1 M
MCF7	16,059,515	10.2 M	3.2 M	0.9 M
T47D	16,719,597	9.9 M	2.8 M	0.8 M



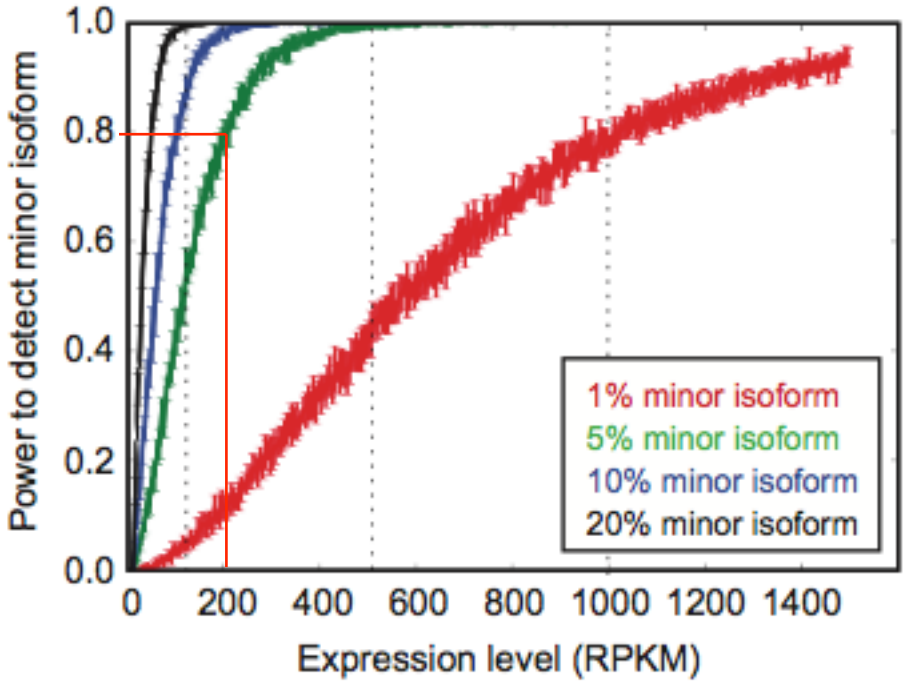
組織特異的なスプライスの例

- 様々な組織由来のサンプルをランすることで、組織特異的なアイソフォームを検出
  - 例）SLC25A3遺伝子では、エクソン3Aと3Bが、相互排他的かつ組織特異的に発現している



どれぐらいの割合で発現しているアイソフォームを検出できるか？

- アイソフォームの発現割合と発現量（RPKM）による検出力の試算
- ある遺伝子に対して、複数のアイソフォーム（スプライスバリアント）が存在
  - そのうちの1つのアイソフォームの割合が5%である場合、RPKMで200カウントとれれば、8割の確率で検出可能と予測



スプライスバリアントの文献 1

- Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimers disease. *PLoS One. 2011 Jan 21;6(1):e16266.*
- Alternative isoform regulation in human tissue transcriptomes. *Nature 456, 470-476 (27 November 2008)*
- Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet. 2008 Dec;40(12):1413-5.*
- Next-generation tag sequencing for cancer gene expression profiling. *Genome Res. 2009 Oct;19(10):1825-35.*
- Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. *Proc Natl Acad Sci U S A. 2010 Mar 16;107(11):5254-9.*
- Characterization of transcriptional complexity during berry development in Vitis vinifera using RNA-Seq. *Plant Physiol. 2010 Apr;152(4):1787-95.*
- Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans. *Nat Struct Mol Biol. 2010 Aug;17(8):1030-4.*

スプライスバリアントの文献 2

- Intron retention facilitates splice variant diversity in calcium-activated big potassium channel populations. *Proc Natl Acad Sci U S A. 2010 Dec 7;107(49):21152-7.*
- RNA-Seq analysis in mutant zebrafish reveals role of U1C protein in alternative splicing regulation. *EMBO J. 2011 May 18;30(10):1965-76.*
- Transcriptome of embryonic and neonatal mouse cortex by high-throughput RNA sequencing. *Proc Natl Acad Sci U S A. 2009 Aug 4;106(31):12741-6.*
- Reaching the depth of the Chinese hamster ovary cell transcriptome. *Biotechnol Bioeng. 2010 Apr 1;105(5):1002-9.*

Data Analysis

- TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics. 2009 May 1;25(9): 1105-11.*
- Towards reliable isoform quantification using RNA-SEQ data. *BMC Bioinformatics. 2010 Apr 29;11*

トランスクリプトームのアセンブル戦略
de novo アセンブル

トランスクリプトームを用いたアセンブルツール Trinity

- Full length transcriptome assembly from RNA-Seq data without a reference genome
  - *Nature Biotechnology* **29**, 644–652 (2011)
- 実験デザイン
  - 酵母およびマウス
  - 75bp x2
  - 約1億リード
- データ解析のながれ
  - 3つのステップにわけて解析

ARTICLES

nature biotechnology

Full-length transcriptome assembly from RNA-Seq data without a reference genome

Manfred G Grabherr[1,8], Brian J Haas[1,8], Moran Yassour[1–3,8], Joshua Z Levin[1], Dawn A Thompson[1], Ido Amit[1], Xian Adiconis[1], Lin Fan[1], Raktima Raychowdhury[1], Qiandong Zeng[1], Zehua Chen[1], Evan Mauceli[1], Nir Hacohen[1], Andreas Gnirke[1], Nicholas Rhind[4], Federica di Palma[1], Bruce W Birren[1], Chad Nusbaum[1], Kerstin Lindblad-Toh[1,5], Nir Friedman[2,6] & Aviv Regev[1,3,7]

Massively parallel sequencing of cDNA has enabled deep and efficient probing of transcriptomes. Current approaches for transcript reconstruction from such data often rely on aligning reads to a reference genome, and are thus unsuitable for samples with a partial or missing reference genome. Here we present the Trinity method for *de novo* assembly of full-length transcripts and evaluate it on samples from fission yeast, mouse and whitefly, whose reference genome is not yet available. By efficiently constructing and analyzing sets of de Bruijn graphs, Trinity fully reconstructs a large fraction of transcripts, including alternatively spliced isoforms and transcripts from recently duplicated genes. Compared with other *de novo* transcriptome assemblers, Trinity recovers more full-length transcripts across a broad range of expression levels, with a sensitivity similar to methods that rely on genome alignments. Our approach provides a unified solution for transcriptome reconstruction in any sample, especially in the absence of a reference genome.

Recent advances in massively parallel cDNA sequencing (RNA-Seq) provide a cost-effective way to obtain large amounts of transcriptome data from many organisms and tissue types[1,2]. In principle, such data can allow us to identify all expressed transcripts[3], as complete and contiguous mRNA sequence from the transcription start site to the transcription end, for multiple alternatively spliced isoforms. However, reconstruction of all full-length transcripts from short reads with considerable sequencing error rates poses substantial computational challenges[4]: (i) some transcripts have low coverage, whereas others are highly expressed; (ii) read coverage may be uneven across the transcript's length, owing to sequencing biases; (iii) reads with sequencing errors derived from a highly expressed transcript may be more abundant than correct reads from a transcript that is not highly expressed; (iv) transcripts encoded by adjacent loci can overlap and thus can be erroneously fused to form a chimeric transcript; (v) data structures need to accommodate multiple transcripts per locus, owing to alternative splicing; and (vi) sequences that are repeated in different genes introduce ambiguity. A successful method should address each challenge, be applicable to both complex mammalian genomes and gene-dense microbial genomes, and be able to reconstruct transcripts of variable sizes, expression levels and protein-coding capacity.

There are two alternative computational strategies for transcriptome reconstruction[4]. Mapping-first approaches[3], such as Scripture[3] and Cufflinks[2], first align all the reads to a reference (unannotated) genome and then merge sequences with overlapping alignment, spanning splice junctions with reads and paired-ends. Assembly-first (*de novo*) methods, such as ABySS[3], SOAPdenovo[6] or Oases (E. Birney, European Bioinformatics Institute, personal communication), use the reads to assemble transcripts directly, which can be mapped subsequently to a reference genome, if available. Mapping-first approaches promise, in principle, maximum sensitivity, but depend on correct read-to-reference alignment, a task that is complicated by splicing, sequencing errors and the lack or incompleteness of many reference genomes. Conversely, assembly-first approaches do not require any read-reference alignments, important when the genomic sequence is not available, is gapped, highly fragmented or substantially altered, as in cancer cells.

Successful mapping-first methods were developed in the past year[4], but substantially less progress was made to date in developing effective assembly-first approaches. As the number of reads grows, it is increasingly difficult to determine which reads should be joined into contiguous sequence contigs. An elegant computational solution is provided by the de Bruijn graph[7,8], the basis for several whole-genome assembly programs[9–11]. In this graph, a node is defined by a sequence of a fixed length of k nucleotides ('k-mer', with k considerably shorter than the read length), and nodes are connected by edges, if they perfectly overlap by $k-1$ nucleotides, and the sequence data support this connection. This compact representation allows for enumerating all possible solutions by which linear sequences can be reconstructed given overlaps of $k-1$.

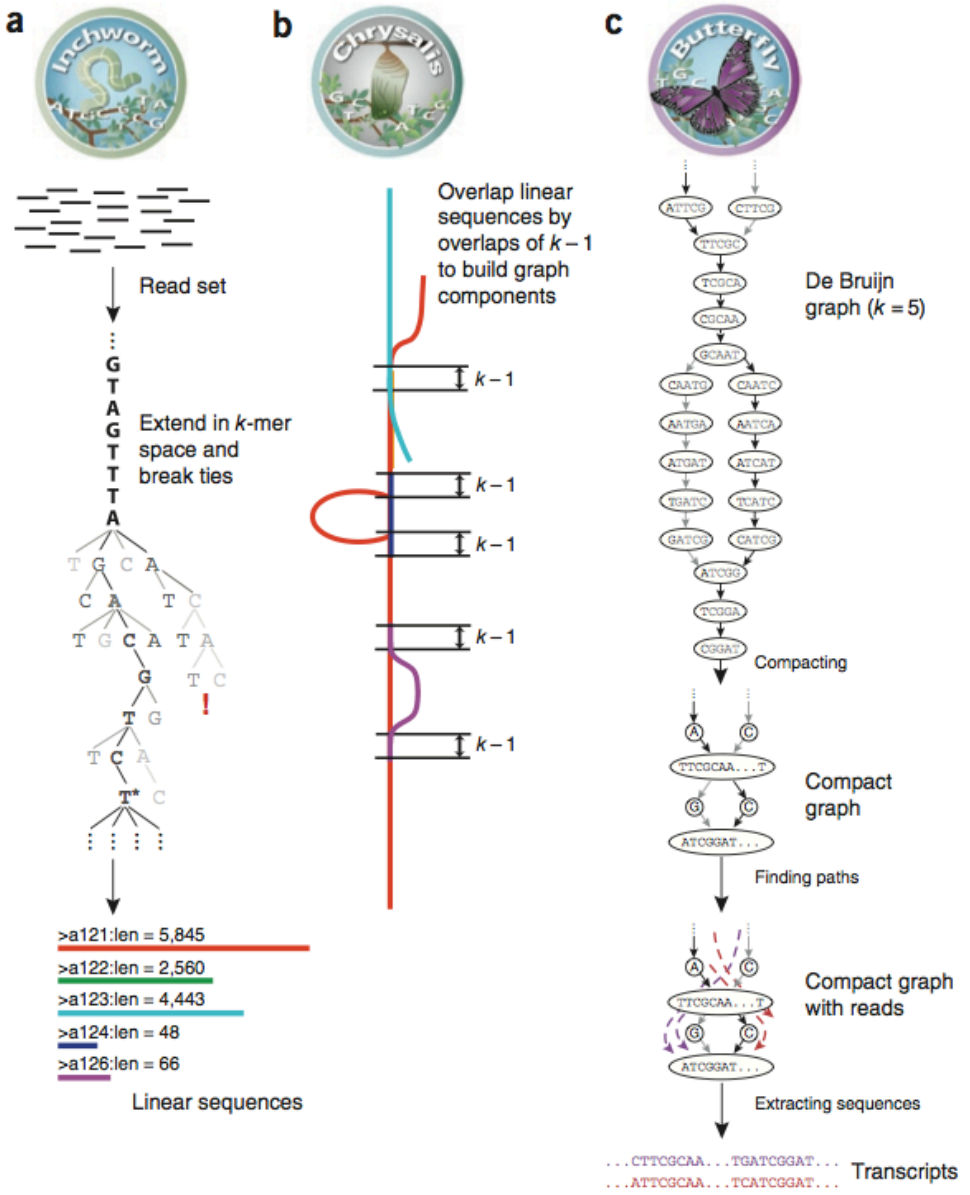
[1]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts, USA. [2]School of Computer Science, Hebrew University, Jerusalem, Israel. [3]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [4]Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts, USA. [5]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden. [6]Alexander Silberman Institute of Life Sciences, Hebrew University, Jerusalem, Israel. [7]Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [8]These authors contributed equally to this work. Correspondence should be addressed to N.F. (nir@cs.huji.ac.il) or A.R. (aregev@broad.mit.edu).

Received 3 December 2010; accepted 28 April 2011; published online 15 May 2011; doi:10.1038/nbt.1883

illumina®

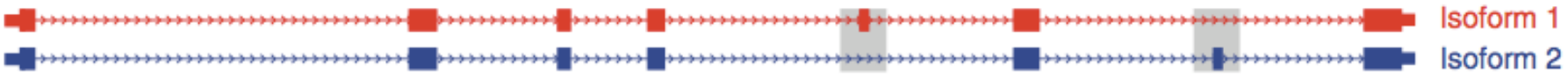
データ解析のながれ

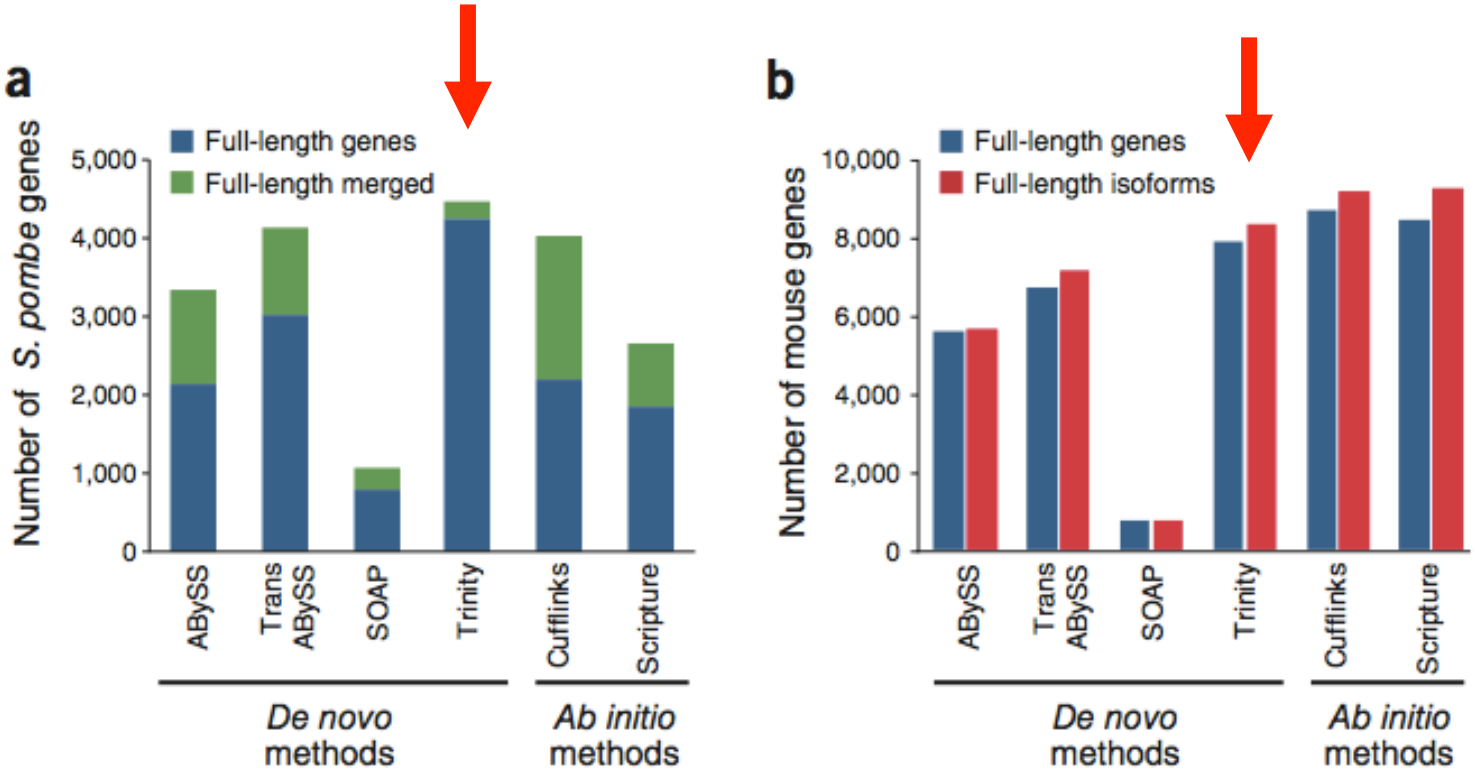
- ステップ1 Inchoworm
  - K-mer を用いてリードをアセンブル
  - K-mer ごとにコンティグを作成
- ステップ2 Chrysalis
  - ステップ1でできたコンティグをプール
    - (k-1)-mer を共有、あるいはコンティグ間ジャンクションにリードがまたがる場合
  - 各プールから de Bruijnグラフを構築
- ステップ3 Butterfly
  - ステップ2でできた de Bruijnグラフを使用しトリムダウンとパスを最小限化
  - リードを使ってグラフを再構成し、各スプライスフォームからひとつの配列を出力



illumina®

スプライスバリアントをも識別、マッピングを利用したソフトと同等の解析結果





Grabherr MG, Haas BJ,Yassour M et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. ***Nat Biotechnol.*** 2011 May 15

illumina®

Trinity アセンブル結果（76bp x2、1億リード）

分裂酵母
(5064遺伝子)

	Scripture (blat)	Cufflinks (blat)	ABySS	Trans-ABySS	SOAP-denovo	Trinity
FL genes	2585	3913	3248	4015	1049	4338
% falsely fused genes	30	45	36	27	26	5
Total contigs	14909	4605	6343	39178	12392	27841
Contigs mapped	11714	3258	4601	31974	5456	7057
Genes captured	3838	4182	4533	4871	3400	4874
Average contig coverage/ gene	4.37	1.07	1.06	5.08	1.01	1.37

マウス

	Scripture (tophat)	Cufflinks (tophat)	ABySS	Trans-ABySS	SOAP-denovo	Trinity
FL transcripts	9086	9010	5561	7025	761	8185
FL genes	8293	8536	5500	6598	760	7749
Total contigs	300148	31121	46783	203085	145518	179340
Contigs mapped	119515	19342	17427	111309	34816	31706
Genes captured	10432	10806	9879	10685	10035	11334
Average contig coverage / gene	12.0	1.65	1.25	5.93	1.12	2.05

トランスクリプトーム de novo の文献 1

- De novo transcriptome assembly with ABySS. *Bioinformatics. 2009 Nov 1;25 (21):2872-7.*
- Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Res. 2010 Oct;20(10):1432-40.*
- De novo assembly and analysis of RNA-seq data. *Nat Methods. 2010 Nov;7(11): 909-12.*
- De novo assembled expressed gene catalog of a fast-growing Eucalyptus tree produced by Illumina mRNA-Seq. *BMC Genomics. 2010 Dec 1;11:681.*
- Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol. 2011 May 15;29(7):644-52.*
- Next-generation transcriptome assembly. *Nat Rev Genet. 2011 Sep 7;12(10): 671-82.*
- Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics. *Mol Biol Evol (2009) 26 (12): 2731-2744.*

トランスクリプトーム de novo の文献 1

- Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology 28, 511–515 (2010)*
- Advancing RNA-Seq analysis. Nat Biotechnol. *2010 May;28(5):421-3.*
- PEACE: Parallel Environment for Assembly and Clustering of Gene Expression. *Nucleic Acids Res. 2010 Jul;38*
- Using deep RNA sequencing for the structural annotation of the Laccaria bicolor mycorrhizal transcriptome. *PLoS One. 2010 Jul 6;5(7):e9780.*
- Scaffolding a Caenorhabditis nematode genome with RNA-seq. *Genome Res. 2010 Dec;20(12):1740-7.*
- De novo sequence assembly and characterization of the floral transcriptome in cross- and self-fertilizing plants. *BMC Genomics 2011, 12:298*
- The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nature Biotechnology 29, 735–741 (2011)*