Most assumptions in molecular biology are wrong

John Mattick
Garvan Institute of Medical Research
ANSTO Distinguished Seminar Series - 08 July 2013
Are we letting a philosophy of the [protein-coding] gene control [our] reasoning? What, then, is the philosophy of the gene? Is it a valid philosophy?

When one starts to question the reasoning behind the origin of the present notion of the gene (held by most geneticists) the opportunity for questioning its validity becomes apparent.

-Barbara McClintock

Letter to Marcus Rhoades, 1950
From The Tangled Field by Nathaniel C. Comfort
The Central Dogma (Crick, 1958) refers to the flow of genetic information from DNA \(\rightarrow\) RNA \(\rightarrow\) protein.

The assumption, based on studies of the *lac* operon in *E. coli*, has been that genes are synonymous with proteins and that most genetic information, including regulatory information, is transacted by proteins.

This protein-centric view reflects a mechanical orientation and has led to several subsidiary assumptions, despite a number of subsequent surprises that should have given pause for thought.

**Surprise #1:** Genes in humans and other complex eukaryotes are mosaics.
**Interpretation:** Introns, despite the fact that they are transcribed, are ‘junk’.

**Surprise #2:** Eukaryote genomes are full of transposon-derived sequences.
**Interpretation:** These sequences are mainly non-functional ‘selfish’ DNA. (!)

**Surprise #3:** Gene number does not scale with developmental complexity.
**Interpretation:** Combinatorial control of transcription, alternative splicing etc. can explain …?
The biggest surprise of the genome projects was the discovery that the number of orthodox (protein-coding) genes does not scale strongly or consistently with complexity:

Humans (and other vertebrates) have approximately the same number of protein-coding genes (~20,000) as C. elegans.

Most of the proteins have similar functions from nematodes to humans, and many are common with brewer’s yeast.

Where is the information that programs our complexity?
The proportion of noncoding DNA broadly increases with developmental complexity.

Irrespective of the extent of non-coding sequences, it is now evident that the vast majority of the genomes of all organisms is transcribed in a dynamic manner in different cells and tissues at different developmental stages.
The Transcriptional Landscape of the Mammalian Genome

The FANTOM Consortium* and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group)*

This study describes comprehensive polling of transcription start and termination sites and analysis of previously unidentified full-length complementary DNAs derived from the mouse genome. We identify the 5′ and 3′ boundaries of 181,047 transcripts with extensive variation in transcripts arising from alternative promoter usage, splicing, and polyadenylation. There are 16,247 new mouse protein-coding transcripts, including 5154 encoding previously unidentified proteins. Genomic mapping of the transcriptome reveals transcriptional forests, with overlapping transcription on both strands, separated by deserts in which few transcripts are observed. The data provide a comprehensive platform for the comparative analysis of mammalian transcriptional regulation in differentiation and development.

Identified > 30,000 transcripts with little or no protein-coding potential
Antisense Transcription in the Mammalian Transcriptome

RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) and the FANTOM Consortium

Antisense transcription (transcription from the opposite strand to a protein-coding or sense strand) has been ascribed roles in gene regulation involving degradation of the corresponding sense transcripts (RNA interference), as well as gene silencing at the chromatin level. Global transcriptome analysis provides evidence that a large proportion of the genome can produce transcripts from both strands, and that antisense transcripts commonly link neighboring "genes" in complex loci into chains of linked transcriptional units. Expression profiling reveals frequent concordant regulation of sense/antisense pairs. We present experimental evidence that perturbation of an antisense RNA can alter the expression of sense messenger RNAs, suggesting that antisense transcription contributes to control of transcriptional outputs in mammals.

~70% of mouse genes exhibit overlapping antisense transcripts
Transcriptional Maps of 10 Human Chromosomes at 5-Nucleotide Resolution


Sites of transcription of polyadenylated and nonpolyadenylated RNAs for 10 human chromosomes were mapped at 5-base pair resolution in eight cell lines. Unannotated, nonpolyadenylated transcripts comprise the major proportion of the transcriptional output of the human genome. Of all transcribed sequences, 19.4, 43.7, and 36.9% were observed to be polyadenylated, non-polyadenylated, and bimorphic, respectively. Half of all transcribed sequences are found only in the nucleus and for the most part are unannotated. Overall, the transcribed portions of the human genome are predominantly composed of interlaced networks of both poly A+ and poly A− annotated transcripts and unannotated transcripts of unknown function. This organization has important implications for interpreting genotype-phenotype associations, regulation of gene expression, and the definition of a gene.

~44% of human transcripts are not polyadenylated and comprise a largely distinct set of sequences
The amazing complexity of the mammalian transcriptome

Graphical representation of the complexity of the transcriptional landscape in mammals. **White boxes** represent non-coding exonic sequences and **blue boxes** protein-coding exonic sequences. **Green diamonds** represent snoRNAs and **orange triangles** represent miRNAs. Indicated are (A) antisense transcripts with overlapping exons, (B) nested transcripts on both strands, (C) antisense transcripts with interlacing exons, and (D) retained introns.

RNA capture-sequencing: focussed transcriptomics a là exome sequencing

• Capture arrays contain probes that hybridize to RNAs expressed from genomic regions of interest

• Transcripts of interest are captured by incubating a RNA sequencing library with the array. Non-target RNAs are washed away.

• Captured transcripts are eluted and enrichment of targeted RNAs confirmed

• Captured transcripts are sequenced
RNA Capture-Seq exposes the deep complexity of the human transcriptome

We used RNA Capture-Seq to examine transcription in intergenic loci that are “gene deserts” as identified by conventional RNA-Seq.

Capture-Seq transforms regions of sparse-mapping RNA-seq reads into long, complex alternatively spliced RNAs (95% of transcripts in intergenic regions completely novel).

Protein-coding loci are similarly transformed, revealing many previously undetected spliced isoforms.

Discovery of 4 new isoforms of p53 by RNA CaptureSeq

FIGURE 2

a

RNAseq

340

b

20,000

Capture Seq

Expression RPKM (x10²)

Expression RPKM (x10²)

p53

393aa

354aa

386aa

393aa

369aa

transactivation

oligomerisation

dna-binding

nuclear local signal

Variant:WT

Domain:

Known alternative exons/splicing

Novel alternative exons/splicing

Known:

Novel:

Exon

Intron

393aa

369aa

386aa

393aa

354aa

Foot

Lung

20,000

Capture Seq

0 5 10 50 100

RNAseq

Expression RPKM (x10²)
Transcriptional and splicing complexity of a noncoding RNA locus revealed by RNA CaptureSeq

Pre-capture:

Post-capture:

AA054531.1

Splice junction frequency

validated by long read sequencing
unvalidated
Table 2. Indices of the functionality of ncRNAs.

<table>
<thead>
<tr>
<th>Feature</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conservation of promoters</td>
<td>[2,27,32]</td>
</tr>
<tr>
<td>Conservation of splice junctions</td>
<td>[27]</td>
</tr>
<tr>
<td>Conservation of sequence</td>
<td>[26,27,32]</td>
</tr>
<tr>
<td>Conservation of genomic position</td>
<td>[31,33,34]</td>
</tr>
<tr>
<td>Conservation of secondary structure</td>
<td>[28–30]</td>
</tr>
<tr>
<td>Positive selection</td>
<td>[230]</td>
</tr>
<tr>
<td>Conservation of expression</td>
<td>[35,36]</td>
</tr>
<tr>
<td>Dynamic expression and alternative splicing</td>
<td>[13,31,32]</td>
</tr>
<tr>
<td>Altered expression or splicing in cancer and other diseases</td>
<td>[37–49]</td>
</tr>
<tr>
<td>Association with particular chromatin signatures</td>
<td>[31,32]</td>
</tr>
<tr>
<td>Regulation by morphogens and transcription factors</td>
<td>[31,32,49,50]</td>
</tr>
<tr>
<td>Tissue- and cell-specific expression patterns</td>
<td>[16,17,19–22,49,51–56]</td>
</tr>
<tr>
<td>Specific subcellular localization</td>
<td>[19–22,52,56]</td>
</tr>
</tbody>
</table>
LncRNA stability compared to mRNAs

<table>
<thead>
<tr>
<th></th>
<th>IncRNAs</th>
<th>mRNAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median half-life</td>
<td>3.5 hr</td>
<td>5.1 hr</td>
</tr>
<tr>
<td>Mean half-life</td>
<td>4.8 hr</td>
<td>7.7 hr</td>
</tr>
</tbody>
</table>

Many IncRNAs are stable
On average, IncRNAs are (slightly) less stable than mRNAs

Clark et al, Genome Research 2012
Non-coding RNA expression in mouse brain

1,328 ncRNAs examined:
849 found to be expressed in brain
60 ubiquitous
623 highly cell- or region-specific

Subcellular localization of long ncRNAs

Mercer, Dinger et al., PNAS 2008
Also called Neat1

MEN ε/β nuclear retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles

Hongjae Sunwoo1, Marcel E. Dinger2, Jeremy E. Wilusz3, Paulo P. Amaral2, John S. Mattick2, and David L. Spector1,4

Abstract

Studies of the transcriptional output of the human and mouse genomes have revealed that there are many more transcripts produced than can be accounted for by predicted protein-coding genes. Using a custom microarray, we have identified 184 non-coding RNAs that exhibit more than 2-fold up- or down-regulation upon differentiation of C2C12 myoblasts into myotubes. Here, we focus on the Men ε/β locus, which is up-regulated 3.3 fold during differentiation. Two non-coding RNA isoforms are produced from a single RNA polymerase II promoter, differing in the location of their 3' ends. Men ε is a 3.2-kb polyadenylated RNA, whereas Men β is a ~20-kb transcript containing a genomically encoded poly(A)-rich tract at its 3' end. The 3' end of Men β is generated by RNase P cleavage. The Men ε/β transcripts are localized to nuclear paraspeckles and directly interact with NONO. Knock-down of MEN ε/β expression results in the disruption of nuclear paraspeckles. Furthermore, the formation of paraspeckles, after release from transcriptional inhibition by DRB treatment, was suppressed in MEN ε/β deleted cells. Our findings indicate that the MEN ε/β non-coding RNAs are essential structural/organizational components of paraspeckles.
The mRNA-like noncoding RNA Gomafu constitutes a novel nuclear domain in a subset of neurons

Masamitsu Sone¹,²,³, Tetsutarō Hayashi², Hiroshi Tarui², Kiyokazu Agata², Masatoshi Takeichi²,³ and Shinichi Nakagawa¹,²,*

¹Nakagawa Initiative Research Unit, RIKEN, 2-1 Hirosawa, Wako 351-0198, Japan
²RIKEN Center for Developmental Biology, 2-2-3 Minatogami Minamimachi, Chuo-ku, Kobe 650-0047, Japan
³Department of Cell and Developmental Biology, Graduate School of Biostudies, Kyoto University, Kitashirakawa, Sakyo-ku, Kyoto 606-8502, Japan

*Author for correspondence (e-mail: nakagawa@riken.jp)

Accepted 21 May 2007
Journal of Cell Science 120, 2498-2506 Published by The Company of Biologists 2007
doi:10.1242/jcs.009357
The long non-coding RNA Gomafu is acutely regulated in response to neuronal activation and involved in schizophrenia-associated alternative splicing

Schizophrenia (SZ) is a complex disease characterized by impaired neuronal functioning. Although defective alternative splicing has been linked to SZ, the molecular mechanisms responsible are unknown. Additionally, there is limited understanding of the early transcriptomic responses to neuronal activation. Here, we profile these transcriptomic responses and show that long non-coding RNAs (lncRNAs) are dynamically regulated by neuronal activation, including acute downregulation of the lncRNA Gomafu, previously implicated in brain and retinal development. Moreover, we demonstrate that Gomafu binds directly to the splicing factors QKI and SRSF1 (serine/arginine-rich splicing factor 1) and dysregulation of Gomafu leads to alternative splicing patterns that resemble those observed in SZ for the archetypal SZ-associated genes DISC1 and ERBB4. Finally, we show that Gomafu is downregulated in post-mortem cortical gray matter from the superior temporal gyrus in SZ. These results functionally link activity-regulated lncRNAs and alternative splicing in neuronal function and suggest that their dysregulation may contribute to neurological disorders.

Molecular Psychiatry advance online publication, 30 April 2013; doi:10.1038/mp.2013.45

Keywords: alternative splicing; Gomafu; neuronal activation; quaking homolog; schizophrenia
Distinct suites of lncRNAs are expressed at different stages of development.

**Genome Research 2008**

Genome-Wide Identification of Long Noncoding RNAs in CD8+ T Cells

Ken C. Pang, Marcel E. Dinger, Tim R. Mercer, Lorenzo Malquori, Sean M. Grimmond, Weisan Chen, and John S. Mattick

**Journal of Immunology 2009**

MEN ε/β nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles

Hongjia Sunwoo, Marcel E. Dinger, Jeremy E. Wilusz, Paulo P. Amaral, John S. Mattick, and David L. Spector

**Genome Research 2009**

Long noncoding RNAs in neuronal-glial fate specification and oligodendrocyte lineage maturation

Tim R. Mercer, Irfan A. Qureshi, Selen Gokhan, Marcel E. Dinger, Guangyu Liu, John S. Mattick, and Mark F. Mehler

**BMC Neuroscience 2009**

SNORD-host RNA Zfas1 is a regulator of mammary development and a potential marker for breast cancer

Marjan E. Askarian-Amiri, Joanna Crawford, Juliet D. French, Chanel E. Smart, Martin A. Smith, Michael B. Clark, Kelin Ru, Tim R. Mercer, Ella R. Thompson, Sunil R. Lakhani, Ana C. Vargas, Ian G. Campbell, Melissa A. Brown, Marcel E. Dinger, and John S. Mattick

**RNA 2011**

The Melanoma-Upregulated Long Noncoding RNA SPRY4-IT1 Modulates Apoptosis and Invasion

Divya Khatana, Marcel E. Dinger, Joseph Maziar, Joanna Crawford, Martin A. Smith, John S. Mattick, and Ranjan J. Perera

**Cancer Research 2011**

Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation


**Embryonic stem cell differentiation**

- Days 2, 4, Blastocyst, Embryoid body

**T-cell activation**

- Memory, Naive

**Myoblast differentiation**

- Neural stem cells, Progenitors, Oligodendrocytes, GABAergic neurons

**Oligodendrogenesis**

**Breast cancer an mammary development**

**Melanoma**
To identify lncRNAs that may be biologically involved in breast cancer, we used NCode ncRNA arrays to analyze RNA from the epithelial cells of developing mouse mammary gland (pregnant, lactating and involuting).

97 lncRNAs were differentially expressed between these developmental stages.
Zfas1 induces mammary epithelial cell proliferation and differentiation

Zfas1 knockdown increases proliferation

Dome formation and beta-casein assay

Zfas1 knockdown increases differentiation of mammary epithelial cells

ME Askarian-Amiri, J Crawford, JD French, CE Smart, MA. Smith, MB Clark, K Ru, TR Mercer, ER Thompson, SR Lakhani, AC Vargas, IG Campbell, MA Brown, ME Dinger and JS Mattick (2011) RNA 17: 878-891.
Hierarchical clustering of NCode expression data across 30 melanoma and 5 normal skin samples revealed hundreds of lncRNAs and mRNAs that are melanoma associated.

We intersected candidates with differentially expressed IncRNAs in melanoma cell lines compared to melanocytes.
**SPRY4-IN1 is an intronic and highly structured IncRNA**

**Genomic context of SPRY4-IN1**

SPRY4-IN1 is expressed from the intron of Sprouty-4 and is conserved only within higher primates.

**SPRY4-IN1 secondary structure**

SPRY4-IN1 has a remarkably stable secondary structure that is conserved within primates.

---

Knockdown of SPRY4-IN1 reduces the invasiveness of melanoma cells

Invasion assay in RNAi treated melanoma cells

The melanoma-upregulated long noncoding RNA SPRY4-IT1 modulates apoptosis and invasion

Effect of SPRY4-IT1 up-regulation on cell motility. Wound healing assay for LOX IMV1 cells transfected with SPRY4-IT1 expressing plasmid and empty vector. Time lapse photography over 12 hours of cells following scratch formation, showing cell migration into the wound.

Differentially expressed noncoding transcripts during embryonic stem cell differentiation

Many lncRNAs have expression profiles that correlate with markers of stem cell differentiation.

Dinger et al. (2008)
Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation.

Differentiation-induced transcripts antisense to developmental genes associate with chromatin modifying complexes and modified histones
Epigenetic processes are central to differentiation and development, long-term responses to environmental variables, and brain function.

Epigenetic memory is embedded in the methylation and hydroxy-methylation of cytosines in DNA and in a wide range of modifications of the histones that package DNA into nucleosomes.

These are catalyzed by a suite of ~60 generic enzymes / chromatin modifying complexes that impose a myriad of different chemical marks at hundreds of thousands, if not millions, of genomic locations in different cells at different stages of differentiation.
Epigenetic processes are central to differentiation and development, long-term responses to environmental variables, and brain function.

Epigenetic memory is embedded in the methylation and hydroxy-methylation of cytosines in DNA and in a wide range of modifications of the histones that package DNA into nucleosomes.

These are catalyzed by a suite of ~60 generic enzymes / chromatin modifying complexes that impose a myriad of different chemical marks at hundreds of thousands, if not millions, of genomic locations in different cells at different stages of differentiation.

What determines the site selectivity of these enzymes?

What determines the positioning of nucleosomes?

What is the molecular basis of epigenome-environmental interactions?
The *Air* Noncoding RNA Epigenetically Silences Transcription by Targeting G9a to Chromatin

Takashi Nagano,1,2 Jennifer A. Mitchell,1 Lionel A. Sanz,3 Florian M. Pauler,4 Anne C. Ferguson-Smith,5 Robert Feil,3 Peter Fraser1,*

*Kcnq1ot1* antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation.

Pandey RR, Mondal T, Mohammad F, Enroth S, Redrup L, Komorowski J, Nagano T, Mancini-Dinardo D, Kanduri C.

Department of Genetics and Pathology, Dag Hammarskjölds Väg 20, Rudbeck Laboratory, Uppsala University, 751 85 Uppsala, Sweden.

Recent investigations have implicated long antisense noncoding RNAs in the epigenetic regulation of chromosomal domains. Here we show that *Kcnq1ot1* is an RNA polymerase II-encoded, 91 kb-long, moderately stable nuclear transcript and that its stability is important for bidirectional silencing of genes in the *Kcnq1* domain. *Kcnq1ot1* interacts with chromatin and with the H3K9- and H3K27-specific histone methyltransferases G9a and the PRC2 complex in a lineage-specific manner. This interaction correlates with the presence of extended regions of chromatin enriched with H3K9me3 and H3K27me3 in the *Kcnq1* domain in placenta, whereas fetal liver lacks both chromatin interactions and heterochromatin structures. In addition, the *Kcnq1* domain is more often found in contact with the nucleolar compartment in placenta than in liver. Taken together, our data describe a mechanism whereby *Kcnq1ot1* establishes lineage-specific transcriptional silencing patterns through recruitment of chromatin remodeling complexes and maintenance of these patterns through subsequent cell divisions occurs via targeting the associated regions to the perinucleolar compartment.
Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression


Contributed by Eric S. Lander, May 3, 2009 (sent for review March 15, 2009)

We recently showed that the mammalian genome encodes >1,000 large intergenic noncoding (linc)RNAs that are clearly conserved across mammals and, thus, functional. Gene expression patterns have implicated these lincRNAs in diverse biological processes, including cell-cycle regulation, immune surveillance, and embryonic stem cell pluripotency. However, the mechanism by which these lincRNAs function is unknown. Here, we expand the catalog of human lincRNAs to ≈3,300 by analyzing chromatin-state maps of various human cell types. Inspired by the observation that the well-characterized lincRNA HOTAIR binds the polycomb repressive complex (PRC2), we tested whether many lincRNAs are physically associated with PRC2. Remarkably, we observe that ≈20% of lincRNAs expressed in various cell types are bound by PRC2, and that additional lincRNAs are bound by other chromatin-modifying complexes. Also, we show that siRNA-mediated depletion of certain lincRNAs associated with PRC2 leads to changes in gene expression, and that the up-regulated genes are enriched for those normally silenced by PRC2. We propose a model in which some lincRNAs guide chromatin-modifying complexes to specific genomic loci to regulate gene expression.

intergenic noncoding (linc)RNAs. These lincRNAs show similar expression levels as protein-coding genes, but lack any protein-coding capacity. Importantly, lincRNAs show significant evolutionary conservation relative to neutral sequences, providing strong evidence that they have been functional in the mammalian lineage (1). We note that nonconserved RNA sequences identified in other collections could be functional, but biological evidence such as loss-of-function experiments would be needed to establish their functionality (5) (Fig. S1A). Previous studies by us and others have demonstrated that groups of lincRNAs exhibit expression patterns across cell types and tissues that correlate with patterns seen for protein-coding genes involved in cellular processes such as cell-cycle regulation, innate immunity responses, and stem cell pluripotency (1, 14). Although these studies clearly demonstrate that there are many functional lincRNAs, key questions remain, including: How many lincRNAs are encoded in mammalian genomes? How do lincRNAs exert their functions? To begin to investigate the number of lincRNAs, we extended our approach of mapping K4-K36 domains to 6 human cell types. The results expand our catalog to
A. Various small RNAs direct chromatin modifications. PIWI proteins and piRNAs interact with HMT/HP1a to induce heterochromatin formation in *Drosophila*. RNA duplexes may be processed in a DICER dependent manner into siRNAs that may subsequently direct chromatin modifications, possibly by targeting nascent transcripts or DNA directly. siRNAs may direct histone methylation (Me) via RITS (RNA-induced transcriptional silencing complex) in centromere heterochromatin in fission yeast.

B. siRNAs originating from RNA Polymerase 4 transcripts can direct DNA methylation in plants.

C. Transcription of SINE B2 elements can establish boundaries between euchromatin and heterochromatin domains in mouse.

D. Long ncRNAs can recruit chromatin repressor complexes (CRC) or chromatin activating complexes (CAC) to target loci in cis or trans, thereby regulating the chromatin context of local genes.

Nucleosomes are preferentially positioned at exons in somatic and germ cells in vertebrates

Regulation of Alternative Splicing by Histone Modifications

Reini F. Luco,1 Qun Pan,2 Kaoru Tominaga,3 Benjamin J. Blencowe,2 Olivia M. Pereira-Smith,3 Tom Mistelit*1

1National Cancer Institute, NIH, Bethesda, MD 20892, USA. 2Banting and Best Department of Medical Research, Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S 3E1, Canada. 3The Barshop Institute for Longevity and Aging Studies, Department of Cellular and Structural Biology, University of Texas Health Science Center, San Antonio, TX 78245-3207, USA.

*To whom correspondence should be addressed. E-mail: mistelit@mail.nih.gov

Alternative splicing of pre-mRNA is a prominent mechanism to generate protein diversity, yet its regulation is poorly understood. We demonstrated a direct role for histone modifications in alternative splicing. We found distinctive histone modification signatures which correlate with splicing outcome in a set of human genes. Modulation of histone modifications causes splice site switching. Histone marks affect splicing outcome by directly modulating the recruitment of splicing regulators via a chromatin-binding protein. These results outline an adaptor system for reading of histone marks by the pre-mRNA splicing machinery.

K4me2, H3-K9ac, H3-K27ac and pan-H4ac were detected (Fig. 1H and fig. S1, D to F). In contrast, H3-K36me3 and H3-K4me1 were enriched over the FGFR2 gene in hMSC, where exon IIIb is repressed, while H3-K27me3, H3-K4me3 and H3-K9me1 were reduced compared to PNT2 cells, where the exon is included (Fig. 1, C to G, and fig. S1, A to C). Histone mark enrichments were not limited to the alternatively spliced exons, but extended along the locus with the highest differences around the alternatively spliced region (Fig. 1 and fig. S1).

Several other PTB-dependent alternatively spliced exons (23) including TPM2 exon 7 and TPM1 exon 3 in hMSC and
Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans


We have recently shown that transcription initiation RNAs (tiRNAs) are derived from sequences immediately downstream of transcription start sites. Here, using cytoplasmic and nuclear small RNA high-throughput sequencing datasets, we report the identification of a second class of nuclear-specific ~17- to 18-nucleotide small RNAs whose 3' ends map precisely to the splice donor site of internal exons in animals. These splice-site RNAs (spliRNAs) are associated with highly expressed genes and show evidence of developmental stage- and region-specific expression. We also show that tiRNAs are localized to the nucleus, are enriched at chromatin marks associated with transcription initiation and possess a 3'-nucleotide bias. Additionally, we find that microRNA-offset RNAs (moRNAs), the miR-15/16 cluster previously linked to oncosuppression and most small nucleolar RNA (snoRNA)-derived small RNAs (sdRNAs) are enriched in the nucleus, whereas most miRNAs and two H/ACA sdRNAs are cytoplasmically enriched. We propose that nuclear-localized tiny RNAs are involved in the epigenetic regulation of gene expression.
Introduction

Animal development and neurological function are critically dependent on inbuilt and environmentally influenced epigenetic processes that alter chromatin structure and hence gene expression patterns at many loci around the genome. Here I consider the implications of the increasing evidence that RNA directs chromatin-modifying complexes to their sites of action, and that RNA is widely edited, especially in the brain. Editing capacity and activity have expanded during vertebrate, mammalian and primate evolution, wherein the majority targets noncoding sequences, many of which are derived from retrotransposed elements. Heuristically joining these dots leads to the obvious possibility that RNA editing alters regulatory circuitry and can feedback into epigenetic memory, and that the expansion of the enzymatic repertoire for RNA editing along with mobilizable target cassettes was central to the emergence of phenotypic plasticity, learning, and cognition. It also suggests that the widespread colonization of mammalian genomes by transposable elements and the pervasive differential transcription of noncoding sequences are not due to selfish elements and noisy transcription, as often thought, but to an evolved capacity that harnessed RNA and retrotransposons as plastic substrates, underpinning phenotypic adaptability and information storage. Finally, the multiple parallels between the nervous and immune systems suggests that they use similar processes, many of which are RNA-related, to induce somatic plasticity and fine scale specificity, especially in intercellular and intermolecular recognition.

Gene-environment Interactions and epigenetic memory

Gene-environment interactions occur at two levels. Short-term responses to physiological variables are largely transduced by signal transduction cascades that alter gene expression.
RNA editing

Two types, both involve base deamination:

A > I - catalyzed by ADARs (“Adenosine Deaminases that Act on RNA”)

ADAR1 and ADAR2 occur in most animals, are expressed in most tissues but highly expressed in brain / nervous system. Developmentally lethal.

ADAR 3 is vertebrate-specific and brain-specific. Function unknown.

C / 5meC > U / T - catalyzed by APOBECs (“ApoB Editing Complex”)

5 families of APOBECs, 3 vertebrate-specific, 2 mammal-specific.

The APOBEC3 family expands from one ortholog in mouse to 8 in human (APOBEC3A-H), with very strong signatures of positive selection.
Table 1: ADAR substrates with editing sites in coding sequence.

<table>
<thead>
<tr>
<th>RNA</th>
<th>Base-pairing</th>
<th>Codon changes[^a]</th>
<th>Functional changes</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glutamate receptor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gluR-B (AMPA)</td>
<td>Exon/intron</td>
<td>Q/R[^607^, R/G[^764^]</td>
<td>Editing of Q/R site lowers Ca[^2+] permeability, &amp; receptor tends to be retained at ER as monomer</td>
<td>[4, 8]</td>
</tr>
<tr>
<td>gluR-D (AMPA)</td>
<td>Exon/intron</td>
<td>R/G[^765^]</td>
<td>Editing of R/G site enhances recovery from desensitization</td>
<td>[8]</td>
</tr>
<tr>
<td>gluR-6 (kainate)</td>
<td>Exon/intron</td>
<td>Q/R[^621^, I/V[^567^], Y/C[^571^]</td>
<td>Editing of all 3 sites in gluR-6 increases higher Ca[^2+] permeability</td>
<td>[4, 7]</td>
</tr>
<tr>
<td>K,1.1 channel</td>
<td>Exon/exon</td>
<td>I/V[^400^]</td>
<td>Rapid recovery from inactivation, shortening duration of and increasing frequency of action potential</td>
<td>[13, 15]</td>
</tr>
<tr>
<td>GABA[^A^]-[^a^]3 receptor</td>
<td>Exon/intron</td>
<td>I/M[^342^]</td>
<td>Smaller peak current amplitudes, slower activation, and faster deactivation compared to non edited receptors</td>
<td>[16, 17]</td>
</tr>
<tr>
<td>Serotonin receptor</td>
<td>Exon/intron</td>
<td>I/V[^157^ &amp; 161^, I/M[^157^, N/D[^159^, N/S[^159^, N/G[^159^</td>
<td>Lower coupling efficacy to G-protein</td>
<td>[20, 21]</td>
</tr>
</tbody>
</table>

[^a] Editing sites are named according to the amino acid change they produced and amino acid position (unedited/edited amino acid position)
Editing is not restricted to a neuroreceptor mRNAs but occurs in thousands of transcripts.

Editing occurs mainly in noncoding sequences, implying that editing is altering regulatory circuits and networks, potentially influencing RNA-directed epigenetic memory.

There is a massive increase in the amount and intensity of A>I editing of human RNAs compared to mouse (35x increase).

The vast majority of this increase occurs in Alu sequences, which are primate-specific SINEs (repetitive sequences) that invaded in three waves during primate evolution and occupy 10.5% of the human genome (~1.2 million largely sequence-unique copies).
The extent and intensity of A>I editing also increases during primate evolution.
Involved in somatic rearrangement and hypermutation of immunoglobulin domains in B-cells and T-cells.

APOBEC3F and 3G appear to control exogenous and endogenous retroviral and LINE-1 retrotransposition. APOBEC3G is expressed in post-mitotic neurons.
L1 retrotransposition in human neural progenitor cells

Nicole G. Coufal¹, José L. Garcia-Perez²,³, Grace E. Peng¹, Gene W. Yeo¹†, Yangling Mu¹, Michael T. Lovci¹†, Maria Morell⁴, K. Sue O'Shea⁴, John V. Moran²,⁵ & Fred H. Gage¹

Long interspersed element 1 (LINE-1 or L1) retrotransposons have markedly affected the human genome. L1s must retrotranspose in the germ line or during early development to ensure their evolutionary success, yet the extent to which this process affects somatic cells is poorly understood. We previously demonstrated that engineered human L1s can retrotranspose in adult rat hippocampus progenitor cells in vitro and in the mouse brain in vivo¹. Here we demonstrate that neural progenitor cells isolated from human fetal brain and derived from human embryonic stem cells support the retrotransposition of engineered human L1s in vitro. Furthermore, we developed a quantitative multiplex polymerase chain reaction that detected an increase in the copy number of endogenous L1s in the hippocampus, and in several regions of adult human brains, when compared to the copy number of endogenous L1s in heart or liver genomic DNAs from the same donor. These data suggest that de novo L1 retrotransposition events may occur in the human brain and, in principle, have the potential to contribute to individual somatic mosaicism.

To determine whether L1 retrotransposition occurred in undifferentiated cells, we conducted immunocytochemical localization of cell-type-restricted markers in EGFP-positive hCNS-SCns. These cells expressed neural stem cell markers, including SOX2, Nestin, Musashi-1 and SOX1 (Fig. 1e and Supplementary Fig. 2a, b), and some co-labelled with Ki-67, indicating that they continued to proliferate (Supplementary Fig. 2c). EGFP-positive hCNS-SCns could also be differentiated to cells of both the neuronal and the glial lineages (Fig. 1f, g). Notably, L1RP did not retrotranspose using our experimental conditions in primary human astrocytes or fibroblasts, although a low level of endogenous L1 expression was detected in both cell types (Fig. 1d and Supplementary Figs 2d, e and 6a, b).

We next used two different protocols to derive NPCs from five human embryonic stem cell lines (hESCs; Fig. 2a). As in our previous study⁴, NPC differentiation led to a ~25-fold increase in L1 promoter activity over a 2-day period, and then a decline (Fig. 2c); there was also a ~250-fold increase in synapsin promoter activity during differentiation (Supplementary Fig. 4b). H13B-derived NPCs expressed both
L1 retrotransposition in neurons is modulated by MeCP2

Alysson R. Muotri*, Maria C. N. Marchetto**, Nicole G. Coufal2, Ruth Oefner², Gene Yeo³, Kinichi Nakashima⁴ & Fred H. Gage²

Long interspersed nuclear elements-1 (LINE-1 or L1s) are abundant retrotransposons that comprise approximately 20% of mammalian genomes¹-³. Active L1 retrotransposons can impact the genome in a variety of ways, creating insertions, deletions, new splice sites or gene expression fine-tuning⁴-⁶. We have shown previously that L1 retrotransposons are capable of mobilization in neuronal progenitor cells from rodents and humans and evidence of massive L1 insertions was observed in adult brain tissues but not in other somatic tissues⁷. In addition, L1 mobility in the adult hippocampus can be influenced by the environment⁸. The neuronal specificity of somatic L1 retrotransposition in neural progenitors is partially due to the transition of a Sox2/HDAC1 repressor complex to a Wnt-mediated T-cell factor/lymphoid enhancer factor (TCF/LEF) transcriptional activator⁹-¹⁰. The transcriptional switch accompanies chromatin remodelling during neuronal differentiation, allowing a transient stimulation of L1 transcription⁷. The activity of L1 retrotransposons during brain development can have an impact on gene expression and neuronal function, thereby increasing brain-specific genetic mosaicism¹¹-¹³. Further understanding of the molecular mechanisms that regulate L1 expression should provide new insights into the role of L1 retrotransposition during brain development. Here we show that L1 neuronal transcription and retrotransposition in rodents are increased in the absence of methyl-CpG-binding protein 2 (MeCP2), a protein involved in global DNA methylation and human neurodevelopmental diseases. Using neuronal progenitor cells derived from human induced pluripotent stem cells and human tissues, we revealed that patients with Rett syndrome (RTT), carrying MeCP2 mutations, have increased susceptibility for L1 retrotransposition. Our data demonstrate that L1 retrotransposition can be controlled in a tissue-specific manner and that disease-related genetic mutations can influence the frequency of neuronal L1 retrotransposition. Our findings add a new level of complexity to the molecular events that can lead to neurological disorders.
Somatic retrotransposition alters the genetic landscape of the human brain

J. Kenneth Baillie1*, Mark W. Barnett1*, Kyle R. Upton1*, Daniel J. Gerhardt2, Todd A. Richmond2, Fioravante De Sapio1, Paul Brennan3, Patrizia Rizzu4, Sarah Smith1, Mark Fell1, Richard T. Talbot1, Stefano Gustinich5, Thomas C. Freeman1, John S. Mattick6, David A. Hume1, Peter Heutink4, Piero Carninci7, Jeffrey A. Jeddeloh2 & Geoffrey J. Faulkner1

Retrotransposons are mobile genetic elements that use a germline ‘copy-and-paste’ mechanism to spread throughout metazoan genomes1. At least 50 per cent of the human genome is derived from retrotransposons, with three active families (L1, Alu and SVA) associated with insertional mutagenesis and disease2–5. Epigenetic and post-transcriptional suppression block retrotransposition in somatic cells6,7, excluding early embryo development and some malignancies6,7. Recent reports of L1 expression8,9 and copy number variation10,11 in the human brain suggest that L1 mobilization may also occur during later development. However, the corresponding integration sites have not been mapped. Here we apply a high-throughput method to identify numerous L1, Alu and SVA germline mutations, as well as 7,743 putative somatic L1 insertions, in the hippocampus and caudate nucleus of three individuals. Surprisingly, we also found 13,692 somatic Alu insertions and 1,350 SVA insertions. Our results demonstrate that retrotransposons mobilize to protein-coding genes differentially expressed and active in the brain. Thus, somatic genome mosaicism driven by retrotransposition may reshape the genetic circuitry that underpins normal and abnormal neurobiological processes.

Mapping the individual retrotransposition events that collectively form a somatic mosaic is challenging owing to the rarity of each mutant allele in a heterogeneous cell population. We therefore developed a high-throughput protocol that we call retrotransposon capture sequencing (RC-seq). First, fragmented genomic DNA was hybridized to custom sequence capture arrays targeting the 3' and 5' termini of full-length L1, Alu and SVA retrotransposons (Fig. 1a and Supplementary Tables 1 and 2). Immobile ERVK and ERV1 long terminal repeat (LTR) elements were included as negative controls. Second, the captured DNA was deeply sequenced, yielding ~25 million paired-end 101-mer reads per sample (Fig. 1b). Last, read pairs were mapped using a conservative computational pipeline designed to identify known (Fig. 1c) and novel (Fig. 1d and Supplementary Fig. 1a–d) retrotransposon insertions with uniquely mapped read pairs (‘diagnostic reads’) spanning their termini.

Previous works have equated L1 CNV with somatic mobilization in vivo10,11. To test this assumption with RC-seq, we first screened five brain subregions taken from three individuals (donors A, B and C) for L1 CNV. A significant (P < 0.001) increase was observed in the number of copies of L1 open reading frame 2 (ORF2) present in DNA
RNA editing appears to be the major mechanism by which environmental signals overwrite encoded genetic information to modify gene function and regulation, particularly in the brain. We suggest that the predominance of Alu elements in the human genome is the result of their evolutionary co-adaptation as a modular substrate for RNA editing, driven by selection for higher-order cognitive function. We show that RNA editing alters transcripts from loci encoding proteins involved in neural cell identity, maturation and function, as well as in DNA repair, implying a role for RNA editing not only in neural transmission and network plasticity but also in brain development, and suggesting that communication of productive changes back to the genome might constitute the molecular basis of long-term memory and higher-order cognition.

The general substrate for A-I editing appears to be double-stranded regions of RNA, but what determines the site selectivity of RNA editing of specific transcripts in different cells and tissues is not well understood [1–3].

RNA editing and gene–environmental interactions in the brain
There are at least three distinct ways that RNA editing can alter brain function in response to experience (i.e. learning) and contribute to the evolution of higher-order cognitive capacities. First, by selectively editing codons and splicing signals in protein-coding sequences involved in modulating fast neurotransmission and all stages of presynaptic vesicle release [1–3], ADAR enzymes can fine-tune the firing properties of neurons required for appropriate neuronal and neural network output and integration. Sec-
REVIEW

RNAs as extracellular signaling molecules

Marcel E Dinger*, Timothy R Mercer* and John S Mattick

Institute for Molecular Bioscience, ARC Special Research Centre for Functional and Applied Genomics, University of Queensland, Brisbane QLD 4072, Australia

(Correspondence should be addressed to J S Mattick; Email: j.mattick@imb.uq.edu.au)

*(M E Dinger and T R Mercer contributed equally to this work)

Abstract

RNA is emerging as a major component of the regulatory circuitry that underpins the development and physiology of complex organisms. Here we review recent evidence that suggests that RNA may supplement endocrine and paracrine signaling by small molecules and proteins, and act as an efficient and evolutionarily flexible source of sequence-specific information transfer between cells, both locally and systemically. As such, RNA signaling may play a central but previously hidden role in multicellular ontogeny, homeostasis, and transmitted epigenetic memory.

Journal of Molecular Endocrinology (2008) 40, 151–159
Paramutation is a heritable epigenetic modification induced in plants by cross-talk between allelic loci. Here we report a similar modification of the mouse Kit gene in the progeny of heterozygotes with the null mutant Kit^{tm1Alf} (a lacZ insertion). In spite of a homozygous wild-type genotype, their offspring maintain, to a variable extent, the white spots characteristic of Kit mutant animals. Efficiently inherited from either male or female parents, the modified phenotype results from a decrease in Kit messenger RNA levels with the accumulation of non-polyadenylated RNA molecules of abnormal sizes. Sustained transcriptional activity at the postmeiotic stages—at which time the gene is normally silent—leads to the accumulation of RNA in spermatozoa. Microinjection into fertilized eggs either of total RNA from Kit^{tm1Alf/+} heterozygotes or of Kit-specific microRNAs induced a heritable white tail phenotype. Our results identify an unexpected mode of epigenetic inheritance associated with the zygotic transfer of RNA molecules.
Transgenerational genetic effects on phenotypic variation and disease risk

Joseph H. Nadeau*

Department of Genetics, Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, OH, USA

Received July 28, 2009; Revised and Accepted July 29, 2009

Traditionally, we understand that individual phenotypes result primarily from inherited genetic variants together with environmental exposures. However, many studies showed that a remarkable variety of factors including environmental agents, parental behaviors, maternal physiology, xenobiotics, nutritional supplements and others lead to epigenetic changes that can be transmitted to subsequent generations without continued exposure. Recent discoveries show transgenerational epistasis and transgenerational genetic effects where genetic factors in one generation affect phenotypes in subsequent generation without inheritance of the genetic variant in the parents. Together these discoveries implicate a key signaling pathway, chromatin remodeling, methylation, RNA editing and microRNA biology. This exceptional mode of inheritance complicates the search for disease genes and represents perhaps an adaptation to transmit useful gene expression profiles from one generation to the next. In this review, I present evidence for these transgenerational genetic effects, identify their common features, propose a heuristic model to guide the search for mechanisms, discuss the implications, and pose questions whose answers will begin to reveal the underlying mechanisms.
The Eukaryotic Genome as an RNA Machine

Paulo P. Amaral, Marcel E. Dinger, Tim R. Mercer, John S. Mattick

The past few years have revealed that the genomes of all studied eukaryotes are almost entirely transcribed, generating an enormous number of non–protein-coding RNAs (ncRNAs). In parallel, it is increasingly evident that many of these RNAs have regulatory functions. Here, we highlight recent advances that illustrate the diversity of ncRNA control of genome dynamics, cell biology, and developmental programming.

RNAs are an integral component of chromosomes and contribute to their structural organization (1, 2). It is now becoming apparent that chromatin architecture and epigenetic memory are regulated by RNA-directed processes that, although the exact mechanisms are yet to be understood, involve the recruitment of histone-modifying complexes and DNA methyltransferases to specific loci (3). Whereas long non-protein-coding RNAs (ncRNAs) have been classically implicated in the regulation of dosage compensation and genomic imprinting in animals (4), they seem to play a much broader role in the epigenetic control of developmental trajectories (5). For example, it was recently shown that 231 long ncRNAs associated with human HOX gene clusters are co-linearly expressed along developmental axes (5), one of which, termed HOTAIR, transcribed from the HOXC locus, was studied in detail and found to recruit Polycomb complexes to repress gene expression in trans at the HOMD cluster (5) (Fig. 1).

Other ncRNAs will likely perform similar functions, such as the intergenic transcripts from globin and antigen receptor loci, which have been associated with complex epigenetic phenomena (6, 7). Small ncRNAs have been consistently linked with heterochromatin formation via the RNA interference (RNAi) pathway (6), including Piwi-interacting RNAs (piRNAs) (9), which guide PIWI family proteins to control transposon activity from flies to vertebrates (10). However, piRNAs might also regulate euchromatin formation, given that PIWI is required for establishing euchromatin in certain subtelomeric regions in Drosophila (11).

Higher-level nuclear organization and chromosomal dynamics are also regulated by ncRNAs in a variety of systems. For example, the formation of the kinetochore and centromeric heterochromatin in fission yeast is dependent on cell cycle-regulated centromeric repeat-derived RNAs and the RNAi pathway, whereas kinetochore assembly and chromosome segregation require the ribonuclease activity of a component of the exosome (12–15). These findings reveal an RNA-based mechanistic link between these processes in mitosis. In Tetrahymena RNA, a direct heterochromatin formation and DNA elimination via RNAi-dependent recruitment of Polycomb complexes and histone methylation (16). The RNAi pathway along with directed histone modifications also regulates the organization of the nucleolus in Drosophila (17).

Likewise, long ncRNAs direct programmed whole-genome rearrangements during clade differential (18). In mammals, transcription of long ncRNAs contributes to various processes including T cell receptor recombination (7), maintenance of telomeres (19, 20), X chromosome pairing required for dosage compensation (21) and inactive X-chromosome perinuclear localization (22).

The functional organization of chromatin can also be regulated by ncRNAs derived from repetitive elements. In mice, bidirectional transcription of a retrotransposed SINE B2 sequence by RNA polymerase (RNAP) II and RNAP III relocates the associated growth hormone locus into nuclear compartments and locally defines the heterochromatin-euchromatin boundary, regulating the expression of the gene during organogenesis (23) (Fig. 1). Given the abundance of transcribed repetitive sequences, this may represent a genome-wide strategy for the control of chromatin domains that may be conserved throughout eukaryotes (23–25). Moreover, each observation and others suggest that a large portion of the genome may, in fact, be functionally active and that transposon-derived sequences may not be reliable indices of the rate of neutral evolution (26).

Transcription

Noncoding RNAs regulate transcription by interacting with transcription factors, RNAP, or...
A simplified biological history of the Earth

- **Multicellular world**
  - Evolution of subcellular structures, and separation of transcription from translation
  - Entry and expansion of introns, evolution of the spliceosome and relaxation of intronic RNA sequences
  - Evolution of RNA-regulatory networks, including the infrastructure for interpreting and acting on these signals (RISC, chromatin-modifying complexes, zinc finger transcription factors, others)

- **Unicellular world**
  - **Archaea**
  - **Eubacteria**
  - **Single-cell eukaryotes** (protista)

- **Expansion of RNA editing**

**Time (mya)**

- -4,000
- -3,000
- -2,000
- -1,000
- Present

**Complexity**

**References**


Institute for Molecular Bioscience
University of Queensland
Marjan Askarian-Amiri, Paulo Amaral, Guy Barry, Pierre Cattenoz, Michael Clark, Joanne Crawford, Larry Croft, Marcel Dinger, Ganqiang Liu, Tim Mercer, Kelin Ru, Martin Smith, Ryan Taft

RIKEN Genomic Science Center
Yoshihide Hayashizaki, Piero Carninci, Harukazu Suzuki, Shintaro Katayama

Allen Brain Institute
Burnham, Florida

Harvard University

Roche / NimbleGen

Johns Hopkins

RIKEN, Wako

AIBN, University of Queensland

Susan Sunkin
Ranjan Perrera
John Rinn
Jeff Jeddeloh
Seth Blackshaw
Shinichi Nakagawa
Ernst Wolvertang

j.mattick@garvan.org.au
The impact of genomics on medicine and healthcare

John Mattick
Garvan Institute of Medical Research
Oxford Nanopore introduces DNA 'strand sequencing' on the high-throughput GridION platform and presents MinION, a sequencer the size of a USB memory stick.

17th February 2012

New generation of sequencing technology uses nanopores to deliver ultra long read length single molecule sequence data, at competitive accuracy, on scalable electronic GridION platform. Miniaturised version of technology, MinION, will make nanopore sequencing universally accessible.

17 February 2012, Oxford, UK/PL, US. Oxford Nanopore Technologies Ltd. today presented for the first time DNA sequence data using its novel nanopore 'strand sequencing' technique and proprietary high performance electronic devices GridION and MinION. These data were presented by Clive G Brown, Chief Technology Officer, who outlined the Company's pathway to a commercial product with highly disruptive features including ultra long read lengths, high throughput on electronic systems and real-time sequencing results. Oxford Nanopore intends to commercialise GridION and MinION directly to customers within 2012.

Oxford Nanopore's GridION system consists of scalable instruments (nodes) used with consumable cartridges that contain proprietary array chips for multi-nanopore sensing. Each GridION node and cartridge is initially designed to deliver tens of Gb of sequence data per 24 hour period, with the user choosing whether to run for minutes or days according to the experiment.
Emerging Molecular Taxonomy
Potentially many low prevalence phenotypes

Slide courtesy of Andrew Biankin
Molecular Taxonomy – Cancer “Biotypes”

- Estrogen Dependent
- Abl-like Kinase Dependant
- EGFR dependent
- HER2 Amplified
- ALK
- Homologous Recombination Defective
- BRAF mutant
- NOVEL mutant

Slide courtesy of Andrew Biankin
High-Throughput Detection of Actionable Genomic Alterations in Clinical Tumor Samples by Targeted, Massively Parallel Sequencing

Nikhil Wagle1,2, Michael F. Berger1, Matthew J. Davis2, Brendan Blumenstiel3, Matthew DeFelice2, Panisa Pochinok1, Matthew Ducar1, Paul Van Hummelen1, Laura E. MacConaill1, William C. Hahn1, Matthew Meyerson1, Stacey B. Gabriel1, and Levi A. Garraway1

ABSTRACT

Knowledge of "actionable" somatic genomic alterations present in each tumor (e.g., point mutations, small insertions/deletions, and copy-number alterations that direct therapeutic options) should facilitate individualized approaches to cancer treatment. However, clinical implementation of systematic genomic profiling has rarely been achieved beyond limited numbers of oncogene point mutations. To address this challenge, we utilized a targeted, massively parallel sequencing approach to detect tumor genomic alterations in formalin-fixed, paraffin-embedded (FFPE) tumor samples. Nearly 400-fold mean sequence coverage was achieved, and single-nucleotide sequence variants, small insertions/deletions, and chromosomal copy-number alterations were detected simultaneously with high accuracy compared with other methods in clinical use. Putatively actionable genomic alterations, including those that predict sensitivity or resistance to established and experimental therapies, were detected in each tumor sample tested. Thus, targeted deep sequencing of clinical tumor material may enable mutation-driven clinical trials and, ultimately, "personalized" cancer treatment.

SIGNIFICANCE: Despite the rapid proliferation of targeted therapeutic agents, systematic methods to profile clinically relevant tumor genomic alterations remain underdeveloped. We describe a sequencing-based approach to identifying genomic alterations in FFPE tumor samples. These studies affirm the feasibility and clinical utility of targeted sequencing in the oncology arena and provide a foundation for genomics-based stratification of cancer patients. Cancer Discovery. 2011;11:82-93. ©2011 AACR.
Whole-genome analysis informs breast cancer response to aromatase inhibition

To correlate the variable clinical features of oestrogen-receptor-positive breast cancer with somatic alterations, we studied pretreatment tumour biopsies accrued from patients in two studies of neoadjuvant aromatase inhibitor therapy by massively parallel sequencing and analysis. Eighteen significantly mutated genes were identified, including five genes (RUNXI, CBF, MYH9, MLL3 and SF3BI) previously linked to haematopoietic disorders. Mutant MAP3K1 was associated with luminal A status, low-grade histology and low proliferation rates, whereas mutant TP53 was associated with the opposite pattern. Moreover, mutant GATA3 correlated with suppression of proliferation upon aromatase inhibitor treatment. Pathway analysis demonstrated that mutations in MAP2K4, a MAP3K1 substrate, produced similar perturbations as MAP3K1 loss. Distinct phenotypes in oestrogen-receptor-positive breast cancer are associated with specific patterns of somatic mutations that map into cellular pathways linked to tumour biology, but most recurrent mutations are relatively infrequent. Prospective clinical trials based on these findings will require comprehensive genome sequencing.
Pancreatic cancer is a highly lethal malignancy with few effective therapies. We performed exome sequencing and copy number analysis to define genomic aberrations in a prospectively accrued clinical cohort \((n = 142)\) of early (stage I and II) sporadic pancreatic ductal adenocarcinoma. Detailed analysis of 99 informative tumours identified substantial heterogeneity with 2,016 non-silent mutations and 1,628 copy-number variations. We define 16 significantly mutated genes, reaffirming known mutations \((\text{KRAS}, \text{TP53}, \text{CDKN2A}, \text{SMAD4}, \text{MLL3}, \text{TGFBRII}, \text{ARID1A} \text{and} \text{SF3BI})\), and uncover novel mutated genes including additional genes involved in chromatin modification \((\text{EPC1} \text{and} \text{ARID2})\), DNA damage repair \((\text{ATM})\) and other mechanisms \((\text{ZIM2}, \text{MAP2K4}, \text{NALCN}, \text{SLC16A4} \text{and} \text{MAGEA6})\). Integrative analysis with \textit{in vitro} functional data and animal models provided supportive evidence for potential roles for these genetic aberrations in carcinogenesis. Pathway-based analysis of recurrently mutated genes recapitulated clustering in core signalling pathways in pancreatic ductal adenocarcinoma, and identified new mutated genes in each pathway. We also identified frequent and diverse somatic aberrations in genes described traditionally as embryonic regulators of axon guidance, particularly \text{SLIT/ROBO} signalling, which was also evident in murine Sleeping Beauty transposon-mediated somatic mutagenesis models of pancreatic cancer, providing further supportive evidence for the potential involvement of axon guidance genes in pancreatic carcinogenesis.
Aussie researchers crack genetic code of deadly pancreatic cancer

NEIL KEENE

A BREAKTHROUGH into how one of the most lethal forms of cancer develops could help save the lives of thousands of people.

The world-first research, headed by Australian scientists, has mapped out the genetic changes that lead to pancreatic cancer, which kills about 2500 people each year in Australia and has one of the highest mortality rates of all cancers.

Garvan Institute of Medical Research Professor Andrew Biankin said a team of more than 100 researchers in Australia, North America and the UK found that pancreatic cancer was not one disease but many, each with its own genetic peculiarities.

The cancer’s variability suggests that traditional one-size-fits-all treatment are likely to be ineffective.

Professor Biankin said the solution could be a more personalised approach, where each patient’s molecular profile could be assessed and their treatment modified accordingly. Just because a cancer comes from the same organ and looks the same under a microscope, they can still be dramatically different, he said.

There might be as many as 100 different sub-types of pancreatic cancer, so it’s no wonder that chemotherapy doesn’t work in all patients.

Getting the treatment right first time is particularly important for pancreatic cancer due to the speed with which it attacks the human body.

About 90 per cent die within a year and only 5 per cent survive more than five years.

Sydney piano teacher Cynthia Terry is one of the lucky few.

She found out she had pancreatic cancer in 2009, but after having her pancreas removed and months of treatment she pulled through.

The tumour from her body was one of more than 100 used during the study to help map out the genetic mutations that lead to the disease.

You go into a very dark place, but I had wonderful doctors, she said.

Cancer survivor Cynthia Terry
Whole-Genome Sequencing for Optimized Patient Management

Matthew N. Bainbridge1,2, Wojciech Wiszniewski3, David R. Murdock1, Jennifer Friedman4,5, Claudia Gonzaga-Jauregui3, Irene Newsham1, Jeffrey G. Reid1, John K. Fink6,7, Margaret B. Morgan1, Marie-Claude Gingras1, Donna M. Muzny1, Linh D. Hoang8, Shahed Yousaf8, James R. Lupski1,3,9,10, and Richard A. Gibbs1,3,*

Abstract

Whole-genome sequencing of patient DNA can facilitate diagnosis of a disease, but its potential for guiding treatment has been under-realized. We interrogated the complete genome sequences of a 14-year-old fraternal twin pair diagnosed with dopa (3,4-dihydroxyphenylalanine)–responsive dystonia (DRD; Mendelian Inheritance in Man #128230). DRD is a genetically heterogeneous and clinically complex movement disorder that is usually treated with l-dopa, a precursor of the neurotransmitter dopamine. Whole-genome sequencing identified compound heterozygous mutations in the SPR gene encoding sepiapterin reductase. Disruption of SPR causes a decrease in tetrahydrobiopterin, a cofactor required for the hydroxylase enzymes that synthesize the neurotransmitters dopamine and serotonin. Supplementation of l-dopa therapy with 5-hydroxytryptophan, a serotonin precursor, resulted in clinical improvements in both twins.
SUMMARY

Personalized medicine is expected to benefit from combining genomic information with regular monitoring of physiological states by multiple high-throughput methods. Here, we present an integrative personal omics profile (iPOP), an analysis that combines genomic, transcriptomic, proteomic, metabolomic, and autoantibody profiles from a single individual over a 14 month period. Our iPOP analysis revealed various medical risks, including type 2 diabetes. It also uncovered extensive, dynamic changes in diverse molecular components and biological pathways across healthy and diseased conditions. Extremely high-coverage genomic and transcriptomic data, which provide the basis of our iPOP, revealed extensive heteroallelic changes during healthy and diseased states and an unexpected RNA editing mechanism. This study demonstrates that longitudinal iPOP can be used to interpret healthy and diseased states by connecting genomic information with additional dynamic omics activity.
Systematic Localization of Common Disease-Associated Variation in Regulatory DNA


Genome-wide association studies have identified many noncoding variants associated with common diseases and traits. We show that these variants are concentrated in regulatory DNA marked by deoxyribonuclease I (DNase I) hypersensitive sites (DHSs). Eighty-eight percent of such DHSs are active during fetal development and are enriched in variants associated with gestational exposure–related phenotypes. We identified distant gene targets for hundreds of variant-containing DHSs that may explain phenotype associations. Disease-associated variants systematically perturb transcription factor recognition sequences, frequently alter allelic chromatin states, and form regulatory networks. We also demonstrated tissue-selective enrichment of more weakly disease-associated variants within DHSs and the de novo identification of pathogenic cell types for Crohn’s disease, multiple sclerosis, and an electrocardiogram trait, without prior knowledge of physiological mechanisms. Our results suggest pervasive involvement of regulatory DNA variation in common human disease and provide pathogenic insights into diverse disorders.
Not too far away... eMedicine

Slide courtesy of Mike Pheasant
The trajectory and opportunity

The accelerating pace of data acquisition is leading to a similar acceleration in the consequent understanding of the molecular basis of disease, genetic predisposition to individual health risk, and the development of new therapies and strategies to reverse or avoid disease.

Within a decade or two, individual genome sequences will be part of everyone’s medical record, and be integrated with other data in mobile electronic records that are both personal and part of larger databases that are used to inform health economics, insurance/underwriting, strategies for reducing disease burdens and costs, and deployment of resources.

Personal genome sequences will likely be done at birth (superseding the present Guthrie test) and at other indicated times throughout life (e.g. to inform cancer diagnosis and treatment, or to assess epigenetic status in complex diseases such as diabetes and neurodegenerative disease.

These genomic analyses will become rebatable after the cost-benefit curves intersect and the medical economy begins to embrace personalised medicine - led by the system and/or demanded by the public.

This information and the databases that sit behind them will lead to a shift to personal and precision medicine, and thence to a system that is primarily geared to health optimisation rather than crisis management.

This will have a progressively greater and ultimately a transformational impact on personal health and wellbeing, health economics, and national productivity. For example, it will identify individuals at risk of many diseases and significantly reduce the incidence of these diseases.

No hospital, let alone individual physician, will have the knowledge to interpret and translate genome sequence into clinically relevant and actionable information. Rather, this advice will be provided online from central sources of evidence-based, well-curated genotype-phenotype correlations, but how these central repositories are built and who controls them is as yet uncertain ... now is the time to start and the attendant opportunity is huge.

Self evidently, it is in the interests of public and private health provision and insurance, and the national interest generally, to embrace these changes and get ahead of the curve.