Diversity through duplication: Whole-genome sequencing reveals novel gene retrocopies in the human population

Sandra R. Richardson[1)], Carmen Salvador-Palomeque[2)] and Geoffrey J. Faulkner[1)3)]*

Gene retrocopies are generated by reverse transcription and genomic integration of mRNA. As such, retrocopies present an important exception to the central dogma of molecular biology, and have substantially impacted the functional landscape of the metazoan genome. While an estimated 8,000–17,000 retrocopies exist in the human genome reference sequence, the extent of variation between individuals in terms of retrocopy content has remained largely unexplored. Three recent studies by Abyzov et al., Ewing et al. and Schrider et al. have exploited 1,000 Genomes Project Consortium data, as well as other sources of whole-genome sequencing data, to uncover novel gene retrocopies. Here, we compare the methods and results of these three studies, highlight the impact of retrocopies in human diversity and genome evolution, and speculate on the potential for somatic gene retrocopies to impact cancer etiology and genetic diversity among individual neurons in the mammalian brain.

Keywords:
■ pseudogene; retrocopy; retrogene; retrotransposition

DOI 10.1002/bies.201300181

[1)] Cancer Biology Program, Mater Medical Research Institute, South Brisbane, QLD, Australia
[2)] Department of Human DNA Variability, Pfizer/University of Granada and Andalusian Regional Government Center for Genomics and Oncology (GENYO), PTS Granada, Granada, Spain
[3)] School of Biomedical Sciences, University of Queensland, Brisbane, QLD, Australia

***Corresponding author:**
Geoffrey J. Faulkner
E-mail: faulknergj@gmail.com

Introduction

Long INterspersed Element 1 (LINE-1 or L1) is a transposable element that has accumulated over time to now account for about one third of human genomic DNA [1]. L1 is classified as an autonomous retrotransposon, as it encodes the enzymatic machinery necessary for its own mobilization via an RNA intermediate. In a round of L1 retrotransposition (Fig. 1A), a donor L1 is transcribed from its original genomic location, and the L1 mRNA translated to give rise to two different proteins. ORF1p is a nucleic acid binding protein [2, 3], and ORF2p harbours endonuclease [4] and reverse transcriptase [5] activities critical for the generation of new insertions. ORF1p and ORF2p exhibit strong cis-preference for binding their encoding mRNA to form the L1 ribonucleoprotein particle (RNP), a hypothesised retrotransposition intermediate [2, 6–10]. The L1 RNP then enters the nucleus, where the ORF2p endonuclease nicks genomic DNA to expose a free 3′ hydroxyl residue, which the ORF2p reverse transcriptase activity uses as a primer to initiate reverse transcription of its associated L1 mRNA, resulting in a new L1 insertion [4, 8, 11]. This process is known as target-site primed reverse transcription (TPRT) [12]. L1 insertions end in a poly-A tail [13] and are frequently flanked by variable length target-site duplications, a hallmark of the retrotransposition process [14].

Although cis-preference is the prevailing rule for template selection by the L1 enzymatic machinery, other transcripts can occasionally be retrotransposed in trans [6, 9, 15–20]. For example, the L1-encoded proteins are responsible for the addition to the human genome of more than one million copies of the nonautonomous retrotransposon *Alu*, and approximately 2,700 copies of the composite retroelement SVA [1]. The L1 enzymatic machinery can also reverse transcribe small RNAs such as the U6 snRNA [21, 22], as well as cellular mRNAs [6, 9, 23] (Fig. 1A).

L1 mobilization of cellular mRNAs is responsible for ~8,000 to 17,000 retrotransposed gene copies identified in the

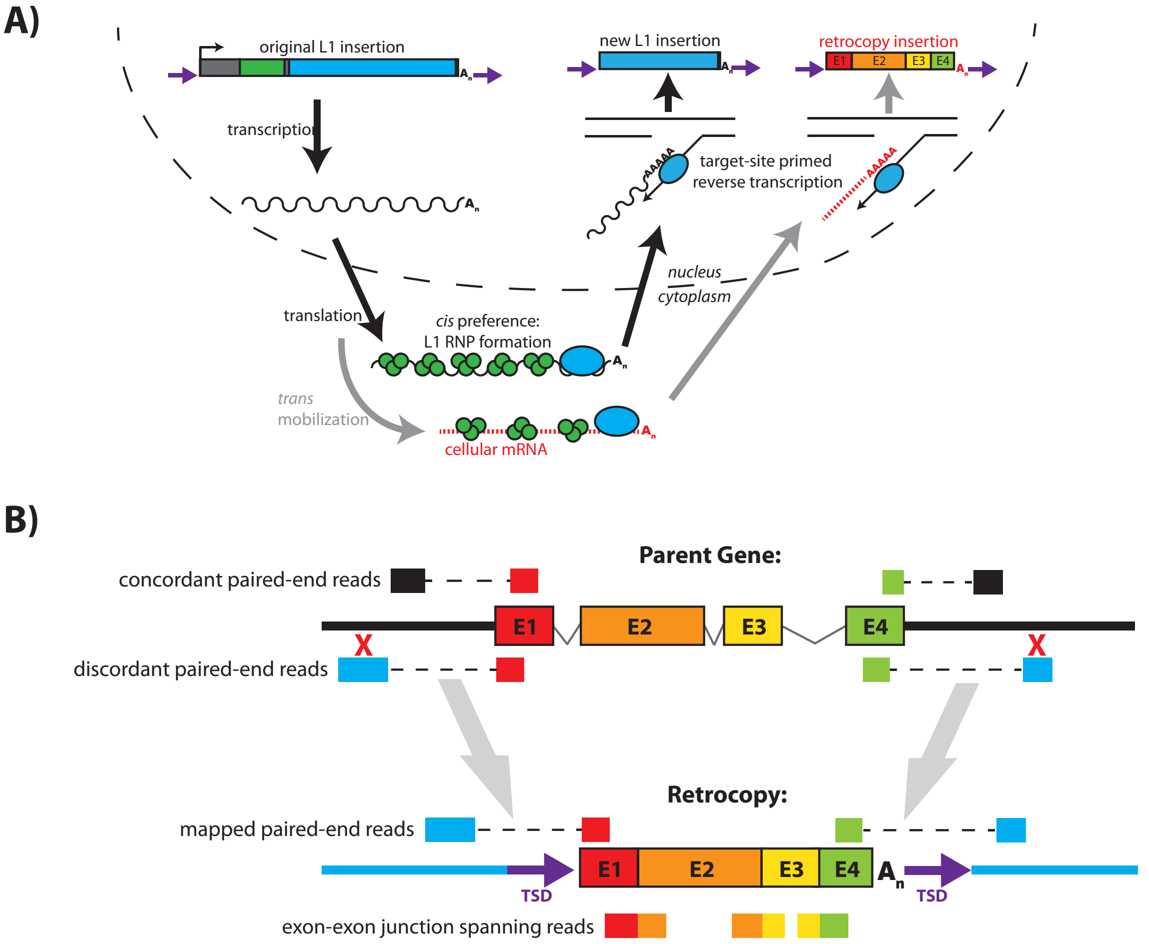

Figure 1. Illustration of retrocopy mechanism and detection strategy. **A:** Generation of L1 insertions and gene retrocopies by the L1 machinery. The typical L1 retrotransposition pathway is indicated by black arrows; gray arrows denote the less-frequent mobilization of cellular mRNAs. Retrotransposition begins with the transcription of a full-length L1 in the genome. The L1 mRNA (wavy black line) is exported from the nucleus and translated, giving rise to the L1 encoded proteins ORF1p (green circles) and ORF2p (blue oval). ORF1p and ORF2p exhibit a strong cis-preference for binding their encoding mRNA, resulting in formation of the L1 ribonucleoprotein particle (RNP). Occasionally, the L1-encoded proteins mobilise cellular mRNAs (dashed red line) in trans. Regardless of the RNA template, insertions generated by the L1-encoded enzymatic machinery undergo target-site primed reverse transcription, resulting in a new L1 copy (blue rectangle), or a gene retrocopy (multi-coloured rectangle), at a distinct genomic location. L1 insertions and retrocopy insertions bear the hallmarks of target-primed reverse transcription, including poly-A tails and target-site duplications (purple arrows). **B:** Characteristics and detection of gene retrocopies. A typical parent gene (above; coloured rectangles denote exons) contains introns (grey lines) and resides at a particular genomic location (heavy black line). Paired-end sequencing reads (dashed lines) wherein one end maps to a gene (red and green rectangles) and the other to its known genomic location (black rectangles) are termed concordant. Conversely, paired-end sequencing reads wherein one end maps to a gene (red and green rectangles), but the other end maps to a distal genomic location (blue rectangles), are termed discordant (denoted by red X's). Discordant paired-end reads are indicative of a gene retrocopy (below), and allow mapping of the retrocopy to its genomic location. Gene retrocopies are distinguished from parent genes by their discrete genomic location (heavy blue line), the presence of retrotransposition hallmarks (target site duplications (TSDs), purple arrows; and a poly-A tail, An) and a lack of introns. Sequencing reads which span exon-exon junctions (bi-coloured rectangles) are also indicative of gene retrocopies.

human genome reference sequence [24–27]. Here, we will refer to any retrotransposed gene copy as a ‘retrocopy’. Retrocopies bear the hallmarks of L1-mediated retrotransposition, including L1 endonuclease preference for the loose consensus sequence 5′-TTTT/AA-3′ [28], flanking target-site duplications, and, as mature mRNAs are mobilised by the L1 machinery, retrocopies lack introns and contain a poly-A tail [29, 30]. A promoter sequence is typically not delivered upon retrotransposition of a cellular mRNA; hence, most retrocopies are not expressed, and presumably lack functionality [29]. These retrocopies are known as processed pseudogenes. Occasionally, a retrocopy may insert into a genomic context permissive of its expression [31–33], giving rise to potentially functional transcripts. We will refer to such an insertion as a ‘retrogene’.

Numerous studies have demonstrated the importance of retrotransposed gene copies in genome evolution and function. Protein-coding retrogenes represent a significant contribution to the ongoing generation of new genes over evolutionary time [34–36]. Gene retrocopies can also influence parent gene expression; for example retrogene-derived siRNAs have been demonstrated to regulate gene expression in mouse oocytes [37, 38] and in the protozoan parasite *T. brucei* [39]. An expressed pseudogene in mouse has also been demonstrated to regulate the mRNA stability of its homologous coding gene [36], and retrogene mRNAs can compete with target gene transcripts for miRNA binding, acting as a molecular 'sink' and interfering with miRNA-mediated gene regulation [40]. Although an instance of a retrocopy insertion associated with human disease has yet to be identified, expression of an *FGF4* retrogene causes short-legged chondrodysplasia in certain breeds of domestic dog [40]. Conversely, retrocopy insertions have the potential to confer protection from infectious disease: in owl monkeys, a retrocopy of the *Cyclophilin A* cDNA into the *TRIM5* gene confers resistance to HIV [32].

Retroelement activity in humans represents an important source of inter-individual genetic variation. Indeed, numerous studies have taken advantage of advances in sequencing technology to uncover novel retrotransposon polymorphisms in the human population (reviewed in [41]). However, the extent of variation between humans with respect to retrocopy insertions is relatively unexplored. To address this issue, three very recent studies by Abyzov et al. [42], Ewing et al. [43], and Schrider et al. [44] analysed whole genome sequencing data generated by the 1,000 Genomes Project Consortium [45], as well as other sources of human genome sequences including The Cancer Genome Atlas (TCGA) [46], to uncover novel polymorphic retrocopies. Below, we compare the retrocopy discovery strategies and results of the three studies, and summarise the biological and evolutionary insights gleaned by each group. We then present our opinions on retrocopy discovery criteria, and speculate regarding the potential impact of germline and somatic retrocopies on inter- and intra-individual human genomic diversity.

Strategies for identifying novel retrocopies

Bona fide retrocopies bear distinguishing features relative to their parent genes: the presence of exon-exon junctions, a genomic address distant from the locus of the parent gene, and the structural hallmarks of L1-mediated retrotransposition. The three studies reviewed here exploited these characteristics to identify and validate novel retrocopies from whole genome sequence data (Table 1). Discordant paired-end reads, in which one end matches exonic sequence of a gene, but the other end aligns to a region of the genome distinct from the known locus of that gene, are indicative of retrocopy insertions, and can be used to map retrocopies to the genome (Fig. 1B). Sequencing reads that span exon-exon junctions are another means to identify retrocopies, since such reads would not arise from intron-containing parent genes (Fig. 1B).

Table 1. Comparison of three recent studies exploiting data from the 1,000 Genomes Project Consortium to uncover novel gene retrocopies

	Human genome sequencing data used	Requirements for calling novel retrocopies	Validation criteria	Novel retrocopies discovered
Abyzov et al. [42]	1,000 Genomes: two deep-sequenced trios, analysed per individual. 968 shallow-sequenced individuals, analysed as pools based on population	Require exon-exon junctions. Calling parameters optimised for each individual or population using a null model based on shifted GENCODE annotations	Read-depth support insertion site found within HuRef genome assembly. PCR validation; DNA sequencing genotyped in additional samples by finding supportive reads	149 Retrocopies absent from human genome reference; 38 with known insertion site. 27 retrocopies present in human genome reference but absent from sequenced genomes
Ewing et al. [43]	1,000 Genomes: 939 shallow-sequenced individuals. Analysed as one pool. The Cancer Genome Atlas (TCGA): 85 paired tumour/ non-tumour genomes	Require insertion site. Require $\geq$8 read pairs spanning retrocopy and insertion location; $\geq$2 read pairs spanning each end of the retrocopy	Precise break points Hallmarks of L1-mediated retrotransposition exon-exon junctions	48 Retrocopies absent from human genome reference (39 present in 1,000 Genomes data, 9 exclusive to TCGA data); 48 with known insertion site. 10 retrocopies present in the human genome reference but absent from sequenced genomes
Schrider et al. [44]	1,000 Genomes: 164 total individuals, including two deep-sequenced trios. Two additional genomes sequenced with SOLiD3 technology	Require insertion site or exon-exon junction. For insertion point, require $\geq$5 paired-end reads spanning retrocopy and insertion location. For exon-exon junctions, require $\geq$1 junction-spanning read with $\geq$10 bp crossing the junction, or $\geq$2 distinct reads with $\geq$5 bp crossing the junction	PCR validation; DNA sequencing genotyped in additional samples by finding supportive reads	73 Retrocopies absent from human genome reference; 21 with known insertion site. 18 retrocopies present in the human genome reference but absent from sequenced genomes

Abyzov et al.

Abyzov et al. employed sequencing data from the 1,000 Genomes Project Consortium, including two deep-sequenced parent-offspring trios and 968 shallow-sequenced individuals representing 14 populations (Table 1). Their strategy identified exon-exon junction spanning reads by comparison to a splice-junction library constructed using GENCODE annotations [47]. Notably, their method did not require insertion point identification to call a novel retrocopy. However, insertion point mapping using discordant paired-end reads was one of three criteria used to support putative retrocopy calls, along with increased read depth in exons, presence of the retrocopy in the HuRef human genome assembly [48], and PCR validation.

Abyzov et al. found 149 novel retrocopies arising from 147 parent genes. Among the 17 retrocopies identified in trios as well as population data, the genomic location of 8 could be identified using discordant paired-end reads, 11 were supported by increased read depth for exons and five present in the HuRef assembly. In addition, 9/17 were supported by splice junction-spanning or exon-insertion site-spanning PCR. Among 132 retrocopies identified only in the population data, a genomic location could be discerned for 28. Insertions discovered in a population were also genotyped in all 14 populations by searching specifically for supportive reads. By this strategy, 72/130 insertions were found in more than one population.

Ewing et al.

Ewing et al. analysed 939 shallow-sequenced genomes representing 13 populations included by the 1,000 Genomes Project Consortium, and 85 paired tumour/non-tumour samples sequenced by TCGA (Table 1). Importantly, the discovery of novel retrocopies relied solely upon discordant paired-end reads. In total, Ewing et al. identified 48 distinct retrocopies from 45 different parent genes, all with known insertion sites. Identification of individual sequence reads containing the retrocopy-genome junction allowed single-nucleotide resolution of breakpoints for 40 retrocopies on at least one end, and for 29 on both ends. Among these 29 insertions, 28 had target-site duplications characteristic L1-mediated mobilization (Fig. 1A). Evidence for exon-exon junctions corresponding to 39/48 insertions was uncovered through local sequence assembly of reads mapping to the region of the putative insertion.

In addition, analysis of TCGA samples revealed evidence for three somatic retrocopy insertions in two lung tumours. Ewing et al. also uncovered novel retrocopies in mouse and chimpanzee genomes. From the deep-sequenced genomes of 17 inbred mouse strains, they found 755 retrocopies derived from 610 parent genes. Among 10 individual chimpanzee genomes, 19 novel retrocopies were identified.

Schrider et al.

Schrider et al. used two retrocopy discovery strategies: mapping of retrocopy insertion sites by discordant paired-end reads, and isolation of sequencing reads containing exon-exon junctions. For the first strategy, Schrider et al. employed sequence data from 15 individuals from the 1,000 Genomes Project Consortium, including two deep-sequenced trios, as well as two genomes they sequenced themselves using the SOLiD3 platform (Table 1). Twenty-one retrocopies with a mapped insertion site were identified in this way, and subsequently genotyped in additional 1,000 Genomes Project Consortium samples. For the second strategy, shallow-sequenced genomes from 149 individuals representing three populations were used to identify unmapped reads corresponding to exon-exon junctions. By this method, 52 novel retrocopies were called based on exon-exon junction data alone.

Interpretations from novel retrocopy discovery

Each of the three studies extended its findings to draw conclusions about L1 biology, human diversity, and genome evolution. Abyzov et al. observed that their set of novel retrocopies was enriched for parent genes that already had known retrocopies present in the human genome reference, leading to the hypothesis that genes involved in specific biological processes may be prone to mobilization by the L1 machinery. Analysis of parent gene functional categories led to the conclusion that genes involved in nuclear envelope breakdown and re-formation, or expressed during the M or M/G1 phase of the cell cycle in HeLa cells, are more likely to be reverse transcribed by the L1 machinery. The authors posited that this correlation indicates a coupling of L1-mediated retrotransposition to cell division, as transcripts highly expressed during cell division would be more likely to undergo retrotransposition. Interestingly, Ewing et al. also observed an enrichment for parent genes that already have retrocopies in the human genome reference, but reported no specific enrichment for cell cycle genes. However, the identification by Ewing et al. of novel retrocopies in tumours is consistent with retrotransposition taking place under conditions of accelerated cell division.

Identifying polymorphic retrocopy insertions can provide insights regarding the contribution of retrocopy insertions to human diversity. Indeed, Abyzov et al. and Ewing et al. found that retrocopy insertions can be population-specific. Furthermore, from 58 segregating retrocopy insertions among 1,025 individuals, Ewing et al. calculated an approximate rate of retrocopy generation at one insertion per 5,177 individuals per generation. Thus, retrocopy insertions contribute to human diversity. Ewing et al. also estimated a rate of retrocopy formation in chimps at one per every 6,804 individuals per generation. A recent study suggested that L1 retrotransposition has differentially influenced human and chimp genome evolution [49]. The ongoing generation of retrocopies by the L1 retrotransposition machinery may likewise contribute to differences in genome evolution between humans and chimps.

Schrider et al. applied their findings to investigate aspects of genome evolution. For example, in support of the hypothesis that natural selection drives gene movement on and off of the X-chromosome [50], they found that fixed functional retrogenes are more likely to have arisen from movement to or from the X-chromosome than novel, polymorphic retrocopies. In addition, their analysis revealed a depletion of fixed retrogenes in introns, while novel retrocopies do not show such a bias. This finding is consistent with the hypothesis that intronic insertions are generally deleterious, whereas intergenic insertions are neutral [31]. Schrider et al. also presented evidence that some of their novel retrocopy insertions can give rise to expressed chimeric transcripts, and that positive selection may act upon retrocopies.

Perspectives, insights and future directions

How important is a mapped insertion site?

Perhaps the most important distinction among discovery strategies is whether identification of the genomic insertion site is required to call a novel retrocopy. Ewing et al. adhered to the most stringent requirements, as their discovery strategy required discordant reads mapping a retrocopy to a genomic locus. Perhaps unsurprisingly, Ewing et al. reported the smallest cohort of novel retrocopies; however, about 56% of their novel retrocopies identified from 1,000 Genomes Project Consortium data were also discovered by at least one of the other two studies. This is a higher rate of overlap than Abyzov et al. (25%) or Schrider et al. (42%), suggesting that calls made based on insertion site may be more reproducible than those made on exon-exon junctions alone. Furthermore, as shown in Fig. 2, among insertions detected by Abyzov et al. or Schrider et al. but not confirmed in another study, the vast majority (99/111 and 38/41, respectively) were called based on exon-exon junctions alone. On the other hand, the datasets generated by Abyzov et al. and Schrider et al. share 17 retrocopies, seven with a mapped insertion site, that were not detected by Ewing et al. Thus, more conservative discovery criteria may result in reduced sensitivity. In general, however, we argue that the 'gold standard' for calling novel retrocopies should entail identification of a genomic location. This criterion is particularly important in cases where the parent gene has one or more known retrocopies in the genome, as reads arising from a known retrocopy could be falsely identified as evidence of a novel retrocopy because of sequencing or alignment errors. Ideally, novel retrocopies should also be demonstrated to bear the hallmarks of L1-mediated retrotransposition.

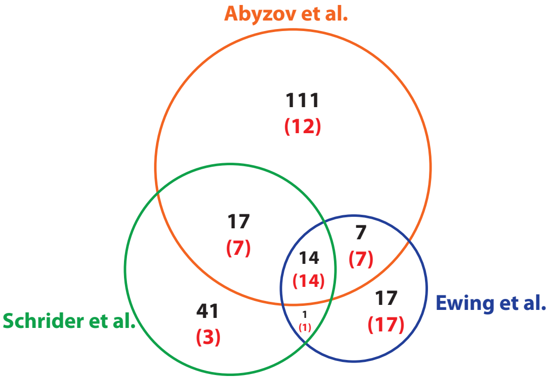

Figure 2. Overlap in retrogene cohorts discovered from 1,000 Genomes Project Consortium data by the three studies. Only retrogenes absent from the human genome reference, but present in 1,000 Genomes Project Consortium data, are represented. Nine retrocopies present only in TCGA data discovered by Ewing et al. are excluded from this comparison. Within each segment, black numbers indicate the total number of novel retrogenes; below, red numbers indicate the number of retrogenes for which an insertion site was mapped by at least one study. For retrocopies with known insertion sites, overlap was confirmed by comparing insertion site coordinates. For those without known insertion sites, only gene names were used. Segments are not drawn to scale.

Implications of novel somatic retrocopies: Cancer and neurons

Although L1-mediated retrotransposition was long considered a germline phenomenon, a growing body of evidence suggests that retrotransposition also occurs in the soma. A number of studies [51–55] have mapped de novo retrotransposon insertions in cancers, and the impact of retrotransposition upon oncogenesis and cancer progression is currently an area of intense study. Furthermore, Ewing et al. discovered three novel gene retrocopy insertions present in lung tumours, but absent from matched normal tissues, demonstrating that retrotransposition of cellular genes can occur in cancer. Retrocopy formation has interesting implications in terms of cancer progression, especially when one considers that oncogenes highly expressed in cancer cells may be likely to undergo retrotransposition and give rise to novel, potentially expressed retrocopies. Alternately, a tumour suppressor gene-derived retrocopy insertion could give rise to antisense transcripts, which could negatively regulate the parent gene through RNA interference mechanisms, contributing to tumourigenesis.

Recent studies have revealed that L1-mediated retrotransposition may occur at high frequency in mammalian neurons [56–59], and in *Drosophila* brain during aging [60]. Whether such events play a physiologically important role, perhaps by effecting differences in gene expression among individual neurons, is a tantalizing question. Indeed, retrotransposition occurs with the highest frequency in the hippocampus [59], the brain region associated with learning and memory. L1-mediated retrotransposition of cellular genes has not yet been detected in neurons, but the potential consequences of new retrocopy insertions in the brain are intriguing. Expressed novel retrocopies of neuron-specific genes could alter gene expression patterns on a cell-to-cell basis, either by increasing levels of already expressed genes, or by negative regulatory mechanisms. It will be interesting to determine whether de novo retrotransposition of cellular transcripts occurs alongside retrotransposon mobilization in the brain, and how such mobilization may alter the functional output of individual neurons.

Retrogene insertions and the timing of retrotransposition events

Although previous analyses indicate that genes highly expressed in a wide variety of tissues are more likely to give rise to novel retrocopies [61], heritable retrotransposition events must take place in the germline (male or female), or in the early embryo prior to germline specification. Notably, there are a large number of known gene retrocopies in the human genome derived from genes active in early embryogenesis and germline development [62–65]. As more genome sequences become available and can be mined for retrocopy data, it will be interesting to observe whether genes with expression patterns restricted to the male or female germline, or the early embryo, have given rise to novel retrocopies. Such events would provide clues about the developmental timing of heritable retrotransposition events.

Conclusions and outlook

As more deep-sequenced genomes become available for analysis, the number of known polymorphic retrocopies in the human population will likely increase. Indeed, high-coverage genomes yield many more novel retrocopies than pooled low-pass genomes: for example in six deeply sequenced individuals, Abyzov et al. detected 17 novel retrocopies (2.8 retrocopies per individual), while analysis of 968 shallowly sequenced genomes yielded only 132 additional retrocopies (0.13 retrocopies per individual). Thus, analysis of low-coverage genomes likely results in a gross underestimation of novel retrocopies in the human population. Furthermore, the three studies reviewed here relied on short, paired-end sequencing reads for mapping putative retrocopies to the human genome reference. PacBio single-molecule real time (SMRT) sequence technology was recently used to sequence complex genomic regions, generating sequence reads of ~1.8 kb average length [66]. Genome sequences generated using this technology will no doubt facilitate confident mapping of transposable element and retrocopy insertions, as individual reads have the potential to span extensive sequence stretches on both sides of an insertion-genome junction.

The three studies highlighted above reveal that, although rare, retrocopies generated by retrotransposition of cellular mRNAs represent an important component of human genetic diversity. Furthermore, their findings provide insight into L1 biology and genome evolution. Indeed, it is interesting to speculate on whether novel retrocopies uncovered by the studies reviewed here will ultimately become fixed functional retrogenes. A 2005 study by Marques et al. estimated a rate of one new functional retrogene per million years in the human lineage [32], suggesting that most novel retrocopies will ultimately be eliminated through negative selection or genetic drift [31]. Generation of retrocopies in somatic cells has important implications for human health and disease, and future studies will no doubt shed light on the consequences of novel retrocopy insertions occurring alongside retroelement insertions in cancer and in the mammalian brain.

Note added in proof, post peer-review

De Boer et al. have reported a male patient afflicted with chronic granulomatous disease resulting from the L1-mediated retrotransposition of a transcript of the TMF1 gene into the X-linked CYBB gene [67]. This is the first documented case of human disease caused by a retrocopy insertion.

Acknowledgements
G.J.F. is supported by an Australian NHMRC Career Development Fellowship (GNT1045237), NHMRC Project Grants GNT1042449, GNT1045991 and GNT1052303, and the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 259743 underpinning the MODHEP consortium.

The authors have declared no conflict of interest.

References

1. **Lander ES, Linton LM, Birren B, Nusbaum C**, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.
2. **Hohjoh H, Singer MF**. 1996. Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. *EMBO J* **15**: 630–9.
3. **Hohjoh H, Singer MF**. 1997. Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. *EMBO J* **16**: 6034–43.
4. **Feng Q, Moran JV, Kazazian HH, Jr, Boeke JD**. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**: 905–16.
5. **Mathias SL, Scott AF, Kazazian HH, Jr, Boeke JD**, et al. 1991. Reverse transcriptase encoded by a human transposable element. *Science* **254**: 1808–10.
6. **Esnault C, Maestre J, Heidmann T**. 2000. Human LINE retrotransposons generate processed pseudogenes. *Nat Genet* **24**: 363–7.
7. **Kulpa DA, Moran JV**. 2005. Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. *Hum Mol Genet* **14**: 3237–48.
8. **Kulpa DA, Moran JV**. 2006. Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. *Nat Struct Mol Biol* **13**: 655–60.
9. **Wei W, Gilbert N, Ooi SL, Lawler JF**, et al. 2001. Human L1 retrotransposition: cis preference versus trans complementation. *Mol Cell Biol* **21**: 1429–39.
10. **Martin SL**. 1991. Ribonucleoprotein particles with LINE-1 RNA in mouse embryonal carcinoma cells. *Mol Cell Biol* **11**: 4804–7.
11. **Cost GJ, Feng Q, Jacquier A, Boeke JD**. 2002. Human L1 element target-primed reverse transcription in vitro. *EMBO J* **21**: 5899–910.
12. **Luan DD, Korman MH, Jakubczak JL, Eickbush TH**. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72**: 595–605.
13. **Scott AF, Schmeckpeper BJ, Abdelrazik M, Comey CT**, et al. 1987. Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. *Genomics* **1**: 113–25.
14. **Grimaldi G, Skowronski J, Singer MF**. 1984. Defining the beginning and end of KpnI family segments. *EMBO J* **3**: 1753–9.
15. **Dewannieux M, Esnault C, Heidmann T**. 2003. LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* **35**: 41–8.
16. **Damert A, Raiz J, Horn AV, Lower J**, et al. 2009. 5′-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. *Genome Res* **19**: 1992–2008.
17. **Hancks DC, Ewing AD, Chen JE, Tokunaga K**, et al. 2009. Exon-trapping mediated by the human retrotransposon SVA. *Genome Res* **19**: 1983–91.
18. **Hancks DC, Goodier JL, Mandal PK, Cheung LE**, et al. 2011. Retrotransposition of marked SVA elements by human L1s in cultured cells. *Hum Mol Genet* **20**: 3386–400.

19. **Ostertag EM, Goodier JL, Zhang Y, Kazazian HH, Jr.** 2003. SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am J Hum Genet* **73**: 1444–51.
20. **Bennett EA, Coleman LE, Tsui C, Pittard WS**, et al. 2004. Natural genetic variation caused by transposable elements in humans. *Genetics* **168**: 933–51.
21. **Buzdin A, Ustyugova S, Gogvadze E, Vinogradova T**, et al. 2002. A new family of chimeric retrotranscripts formed by a full copy of U6 small nuclear RNA fused to the 3′ terminus of L1. *Genomics* **80**: 402–6.
22. **Garcia-Perez JL, Doucet AJ, Bucheton A, Moran JV**, et al. 2007. Distinct mechanisms for trans-mediated mobilization of cellular RNAs by the LINE-1 reverse transcriptase. *Genome Res* **17**: 602–11.
23. **Mandal PK, Ewing AD, Hancks DC, Kazazian HH, Jr.** 2013. Enrichment of processed pseudogene transcripts in L1-ribonucleoprotein particles. *Hum Mol Genet* **22**: 3730–48.
24. **Potrzebowski L, Vinckenbosch N, Kaessmann H**. 2010. The emergence of new genes on the young therian X. *Trends Genet* **26**: 1–4.
25. **Henrichsen CN, Vinckenbosch N, Zollner S, Chaignat E**, et al. 2009. Segmental copy number variation shapes tissue transcriptomes. *Nat Genet* **41**: 424–9.
26. **Marques AC, Vinckenbosch N, Brawand D, Kaessmann H**. 2008. Functional diversification of duplicate genes through subcellular adaptation of encoded proteins. *Genome Biol* **9**: R54.
27. **Potrzebowski L, Vinckenbosch N, Marques AC, Chalmel F**, et al. 2008. Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biol* **6**: e80.
28. **Jurka J**. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci USA* **94**: 1872–7.
29. **Vanin EF**. 1985. Processed pseudogenes: characteristics and evolution. *Annu Rev Genet* **19**: 253–72.
30. **Weiner AM, Deininger PL, Efstratiadis A**. 1986. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annu Rev Biochem* **55**: 631–61.
31. **Vinckenbosch N, Dupanloup I, Kaessmann H**. 2006. Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci USA* **103**: 3220–5.
32. **Marques AC, Dupanloup I, Vinckenbosch N, Reymond A**, et al. 2005. Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol* **3**: e357.
33. **Vinckenbosch E, Robichon F, Eliez S**. 2005. Gray matter alteration in dyslexia: converging evidence from volumetric and voxel-by-voxel MRI analyses. *Neuropsychologia* **43**: 324–31.
34. **Chen S, Krinsky BH, Long M**. 2013. New genes as drivers of phenotypic evolution. *Nat Rev Genet* **14**: 645–60.
35. **Long M, VanKuren NW, Chen S, Vibranovski MD**. 2013. New gene evolution: little did we know. *Annu Rev Genet* **47**: 307–33.
36. **Kaessmann H, Vinckenbosch N, Long M**. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* **10**: 19–31.
37. **Watanabe T, Totoki Y, Toyoda A, Kaneda M**, et al. 2008. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453**: 539–43.
38. **Valley JF, Guerid A, Lerch P, Pache G**, et al. 1979. Application of the thermoluminescent dosimetry to a pion beam. *Radiat Environ Biophys* **16**: 225–9.
39. **Burki F, Kaessmann H**. 2004. Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. *Nat Genet* **36**: 1061–3.
40. **Betran E, Emerson JJ, Kaessmann H, Long M**. 2004. Sex chromosomes and male functions: where do new genes go? *Cell Cycle* **3**: 873–5.
41. **Beck CR, Garcia-Perez JL, Badge RM, Moran JV**. 2011. LINE-1 elements in structural variation and disease. *Annu Rev Genomics Hum Genet* **12**: 187–215.
42. **Abyzov A, Iskow R, Gokcumen O, Radke DW**, et al. 2013. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. *Genome Res* **23**: 2042–52.
43. **Ewing AD, Ballinger TJ, Earl D, Harris CC**, et al. 2013. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome Biol* **14**: R22.
44. **Schrider DR, Navarro FC, Galante PA, Parmigiani RB**, et al. 2013. Gene copy-number polymorphism caused by retrotransposition in humans. *PLoS Genet* **9**: e1003242.
45. **Abecasis GR, Altshuler D, Auton A, Brooks LD**, et al. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–73.
46. **Weinstein JN, Collisson EA, Mills GB, Shaw KR**, et al. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**: 1113–20.
47. **Harrow J, Denoeud F, Frankish A, Reymond A**, et al. 2006. GENCODE: producing a reference annotation for ENCODE. *Genome Biol* **7 Suppl 1**: S4.1–9.
48. **Axelrod N, Lin Y, Ng PC, Stockwell TB**, et al. 2009. The HuRef Browser: a web resource for individual human genomics. *Nucleic Acids Res* **37**: D1018–24.
49. **Marchetto MC, Narvaiza I, Denli AM, Benner C**, et al. 2013. Differential L1 regulation in pluripotent stem cells of humans and apes. *Nature* **503**: 525–9.
50. **Emerson JJ, Kaessmann H, Betran E, Long M**. 2004. Extensive gene traffic on the mammalian X chromosome. *Science* **303**: 537–40.
51. **Miki Y, Nishisho I, Horii A, Miyoshi Y**, et al. 1992. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res* **52**: 643–5.
52. **Iskow RC, McCabe MT, Mills RE, Torene S**, et al. 2010. Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* **141**: 1253–61.
53. **Lee E, Iskow R, Yang L, Gokcumen O**, et al. 2012. Landscape of somatic retrotransposition in human cancers. *Science* **337**: 967–71.
54. **Solyom S, Ewing AD, Rahrmann EP, Doucet T**, et al. 2012. Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Res* **22**: 2328–38.
55. **Shukla R, Upton KR, Munoz-Lopez M, Gerhardt DJ**, et al. 2013. Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell* **153**: 101–11.
56. **Muotri AR, Chu VT, Marchetto MC, Deng W**, et al. 2005. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**: 903–10.
57. **Muotri AR, Marchetto MC, Coufal NG, Oefner R**, et al. 2010. L1 retrotransposition in neurons is modulated by MeCP2. *Nature* **468**: 443–6.
58. **Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW**, et al. 2009. L1 retrotransposition in human neural progenitor cells. *Nature* **460**: 1127–31.
59. **Baillie JK, Barnett MW, Upton KR, Gerhardt DJ**, et al. 2011. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**: 534–7.
60. **Li W, Prazak L, Chatterjee N, Gruninger S**, et al. 2013. Activation of transposable elements during aging and neuronal decline in Drosophila. *Nat Neurosci* **16**: 529–31.
61. **Yi X, Liang Y, Huerta-Sanchez E, Jin X**, et al. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**: 75–8.
62. **Booth HA, Holland PW**. 2004. Eleven daughters of NANOG. *Genomics* **84**: 229–38.
63. **Elliman SJ, Wu I, Kemp DM**. 2006. Adult tissue-specific expression of a Dppa3-derived retrogene represents a postnatal transcript of pluripotent cell origin. *J Biol Chem* **281**: 16–9.
64. **Liedtke S, Enczmann J, Waclawczyk S, Wernet P**, et al. 2007. Oct4 and its pseudogenes confuse stem cell research. *Cell Stem Cell* **1**: 364–6.
65. **Pain D, Chirn GW, Strassel C, Kemp DM**. 2005. Multiple retropseudogenes from pluripotent cell-specific gene expression indicates a potential signature for novel gene identification. *J Biol Chem* **280**: 6265–8.
66. **Huddleston J, Ranade S, Malig M, Antonacci F**, et al. 2014. Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res*, in press, doi: 10.1101/gr.168450.113.
67. **de Boer M, van Leeuwen K, Geissler J, Weemaes CM**, et al. 2014. Primary immunodeficiency caused by an exonized retroposed gene copy inserted in the CYBB gene. *Hum Mutat*, in press, doi: 10.1002/humu.22519.