How to deduplicate PCR

Vivien Marx

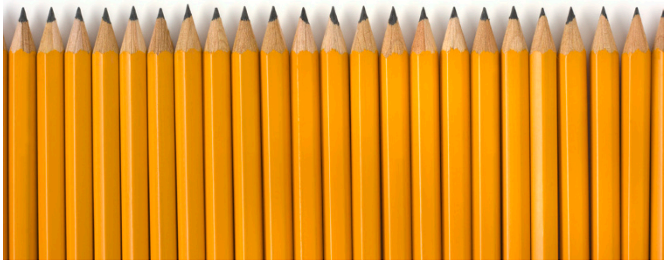
PCR duplicates—sequencing reads from the same original genomic fragment—can cause headaches. But there are remedies.

Plenty can go wrong when genomic material is amplified. The resulting fragment can be less than 100% identical to its DNA or RNA template. The polymerase chain reaction (PCR) can also selectively amplify one fragment over another "despite our best efforts," says Marie Adams, who directs the genomics core at the Van Andel Research Institute. And labs must watch out for PCR duplicates, which occur routinely when the same DNA or RNA fragment is amplified and sequenced multiple times.

Such identical reads "waste space" on a sequencer flow cell, says Adams, and use resources that could have been used for potentially informative reads. When PCR duplicates obstruct an accurate count of DNA or RNA molecules, they can bias many types of high-throughput-sequencing experiments—those involving human genes or model organism genes, a microbiome or immune cell responses.

Gary Schroth, distinguished scientist at Illumina, remembers a worried researcher who confessed to him that out of fear of PCR bias, the person routinely prepared sequencing libraries with fewer PCR cycles than the company recommended. Even though PCR paranoia is understandable, labs need not give into it in all instances. Schroth and his Illumina team have found that PCR issues have a "very, very minimal" effect in a typical RNA-sequencing (RNA-seq) experiment. In sequencing runs to find germline mutations, PCR duplicates can be identified and excluded because shearing creates fragments that have unique start and stop positions.

Fragmentation patterns also help scientists deal with PCR duplicates in shallow genome sequencing to identify copy-number changes, especially in paired-end sequencing, says Kasper Karlsson, a postdoctoral fellow at Stanford University School of Medicine.



Y. Yang/Getty Images

PCR duplicates can ruin an experiment. They occur when the same genomic fragment is amplified and sequenced multiple times.

Similarly, labs can use fragmentation patterns when profiling immune-cell receptors with high-throughput transcriptome sequencing, and "no molecular barcoding is needed," says Dmitriy Chudakov, a researcher who splits his time between the Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry in Moscow and Masaryk University in Brno, Czech Republic. In the "absolute majority" of cases, he says, PCR duplicates can be safely removed during data analysis.

But for accurate quantification, molecular barcodes are needed. In targeted resequencing, all or many reads can look similar, and a researcher will not know "the real numbers of incoming molecules," says Chudakov. He hopes that in the near future, all resequencing experiments will use unique molecular identifiers (UMIs) to enable accurate counting and tracking of molecules.

UMI strategies

UMIs can help labs track molecules and remove errors in amplification and sequencing[1–7]. Every molecule in a sample, be it a genomic fragment or cDNA, is uniquely labeled with an UMI, typically a random oligonucleotide sequence, before the PCR step. When analysis reveals two identical tags on two identical sequences, they are from the same original molecule, says Ian Sudbery, bioinformatician at the University of Sheffield. A finding of two different tags on the same sequence means two different original molecules. The reality is trickier, but this is the basic principle.

UMIs are widely used, especially for single-cell analysis, where they have solved the "severe quantitative bias introduced by PCR," says Sten Linnarsson, a researcher at Karolinska Institute who develops methods for quantitative single-cell RNA-sequencing (scRNA-seq). UMIs can help scientists correct base-substitution errors that occur during DNA sequencing. "They are a workhorse of single-cell genomics, easy to use and widely deployed," he says.

To ensure accuracy, PCR duplicates need to be removed in scRNA-seq, especially with transcripts expressed at low levels, all of which explains why scRNA-seq labs were early UMI adopters, says Karlsson. Unlike amplicon sequencing, with input molecules aplenty and thousands of genome copies, single-cell-oriented work involves tiny amounts of starting material and therefore many amplification rounds. Karlsson wrote his dissertation on UMIs while in Linnarsson's lab and is currently developing a quantitative way to assess tumor evolution by monitoring barcoded subclones.

When fragmentation is a post-PCR step, such as in several single-cell sequencing protocols, UMIs matter, says Sudbery. And UMIs can be useful in experiments in which

E. van Bodegon/Getty Images

Unique identifiers can help labs sort true sequencing reads from those with amplification or sequencing errors.

DNA or RNA fragmentation isn't random, such as those involving 4C techniques in which fragment analysis helps researchers discern chromosome configuration, and iCLIP experiments to identify RNA–protein binding sites. In his view, UMIs might be needed in traditional RNA-seq, he says, because there are a large number of possible fragmentation patterns but an even larger number of original molecules.

Molecular identifiers are always useful, says Adams, but labs have to consider when they are most worth it. In a one-off bulk RNA-seq experiment that looks only at expression data, "it's probably not worth the extra work and cost," she says. When a lab does a more sensitive experiment with scRNA-seq or total RNA-seq with the intent of more nuanced counting, the benefits of molecular identifiers are more salient, she says. Such experiments include isoform analysis, large-scale translational or clinical sequencing projects.

Designing UMIs to have a specified sequence is pricy, whereas UMIs with a more randomized sequence are almost negligible in cost, says Karlsson. For an absolute molecule count and when exhaustive sequencing of an UMI-labeled sample is the plan, a researcher will want to see most UMIs at least twice, even thrice for the highest confidence. To achieve that, and given the uneven way PCR amplifies material, a lab would need to sequence around 10–20 reads per UMI on average.

Singletons—UMI tags supported by a single read—can be an issue. Even in deep sequencing, singletons are common, says Karlsson, and they tend to be due to PCR or sequencing errors, which makes them readily removable from the analysis. But an extreme scenario is possible: only one read maps to a specific spot in the genome. Such a singleton is one researchers should keep. Another type of confounder is collision: by chance two templates and their associated UMI sequences are identical, but they stem from separate molecules, not ones amplified from the same template.

UMIs are a balancing act: a low-complexity tag will increase the chances of UMI collision, and a high-complexity tag—for example, one with a long, randomized sequence—will increase the chance of PCR or sequencing errors in the tag and complicate data analysis, says Karlsson. A very long tag can also form secondary structures, which can skew the PCR reaction.

Coping with UMI errors

Molecular barcodes can correct sequencing errors, but they themselves can fall victim to amplification or sequencing errors. The result can be a molecular miscount. As Chudakov explains, coping with UMI errors takes forethought. For example, most errors, including those in UMIs, happen at late stages of amplification. If a lab has a low number of molecules and amplifies a lot, he says, "you will get high coverage in terms of sequencing reads per UMI." He recommends setting a threshold of around ten reads per UMI to correct errors. UMIs that have one or two mismatches can then be either discarded or clustered, says Chudakov. There are many options, but, he says, it takes careful use of software-based analysis.

Good design of molecular identifiers can ease software-based analysis. One common bit of wisdom, says Adams, is to design identifiers to be at least three base pairs from one another in 'sequence space'. In an experiment, if up to two bases are altered as a result of PCR errors, the molecular identifier will still look more like its parent label than like any other molecular identifier in the pool. But this approach means that researchers need fairly long molecular identifiers, with 10–14 base pairs, in order to create a diverse library pool. "But you don't want to make them too long—if they are, then you're using valuable sequencing reads on these 'nonbiological' bases," she says. Picking too few biological base pairs lowers the rate with which sequences can be mapped back to the genome.

Tools to compute

Marking of PCR duplicates is usually done with software tools such as Picard's MarkDuplicates, developed at the Broad Institute of Harvard and MIT, or samtools rmdup, developed at the Sanger Institute, says Benjamin Johnson, a researcher in Van Andel's bioinformatics core. Both tools identify aligned fragments that have the same genome coordinates, which suggests that the duplicates came from the same original fragment and are therefore technical artifacts. The two tools use different but related criteria to decide which duplicates to keep in the analysis.

As single-cell applications such as scRNA-seq have emerged, labs have sought new ways to identify the original number of transcripts and remove the technical duplicates. UMIs are helpful in the way they 'collapse' duplicate reads to a single unique fragment. MarkDuplicates now has expanded functionality so users can include molecular barcodes. Some commercial tools involve a deduplication step; the computational pipelines offered by 10X Genomics are one example. Among academically developed software, Johnson especially likes UMI-tools, developed by the Sudbery lab and the computational genomics analysis and training group at the MRC Weatherall Institute of Molecular Medicine[8].

The motivation for the software stems from the researchers' experience with data from iCLIP experiments to identify the binding sites of proteins on RNA. When viewing all the reads that mapped to the same location, one would expect the UMIs associated with this group to differ from one another by a certain number of bases, says Sudbery, "but when we looked at our data, for many of our binding sites, all of the UMIs were only one base different from each other, which didn't seem right."

University of Sheffield

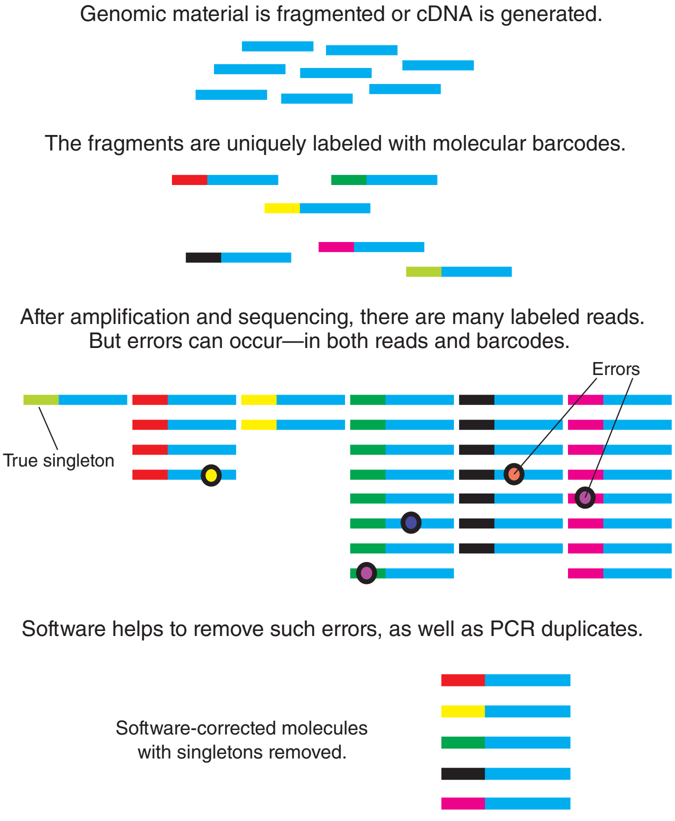
Scientists usually assume UMIs are error-free, says Ian Sudbery, but they might not be.

PCR error or sequencer error can indeed alter the UMI tag, says Sudbery, and deeper sequencing only makes things worse because it propagates the errors through the results. What worsens matters further is sequencing that is very deep compared to the complexity of the library being sequenced, such as with iCLIP experiments, he says. Presently, scRNA-seq experiments do not involve deep sequencing, so the problem is less extreme. "However, I suspect this will change in time as sequencing gets cheaper," he says. "Even so, we do see changes in the clustering of samples from single-cell RNA-seq after applying our method for dealing with this problem."

Scientists usually assume UMIs are error-free, says Sudbery, and the challenge with UMI errors would benefit from a comprehensive analysis to model the mathematics of PCR duplicate occurrence. But a number of labs have developed tools to handle UMI errors in data analysis. The software he codeveloped, UMI-tools, looks at groups of reads that map to the same genomic location or reads that come from the same transcript and assesses the UMI sequences of those reads. Starting with the UMI sequence that has been observed most often at a given site, the software asks which of the other sequences could have arisen as an error that changed it from the original. One UMI may have arisen via an error from another if its sequence differs by only one or several bases and if it is seen less than half as often as the other.

The software evaluates each UMI iteratively. The UMI sequences are grouped and UMIs are selected that could not have arisen as duplicates of any of the others, and these are then chosen as "representatives of this group," says Sudbery. The software returns to the list of UMIs that map to the same position or transcript, and the process is repeated until all UMIs are dealt with. Most users accessing the software appear to be using it for scRNA-seq, he says. In his lab, MarkDuplicates is used to handle data without UMIs. The software is part of most variant-calling and ChIP-seq pipelines, although in general it is not used in RNA-seq because, in his view, fragmentation patterns are not a reliable way to detect duplicates in these data. For labs using randomly generated UMIs, he recommends UMI-tools, as MarkDuplicates does not, in his view, consider the possibility of UMI-based sequencing or PCR errors. "This is the niche that our tool fills," he says.

The 'real' solution with UMIs, says Sudbery, is not to use random barcodes, but to synthesize a pool of thousands of oligo sequences designed such that a PCR or sequencing error would not create sequence similarity. Some companies, such as Bioo Scientific, offer oligo kits, "but of course you have to pay for this," he says.


Genomic material is fragmented or cDNA is generated.
The fragments are uniquely labeled with molecular barcodes.
After amplification and sequencing, there are many labeled reads.
But errors can occur—in both reads and barcodes.
Errors
True singleton
Software helps to remove such errors, as well as PCR duplicates.
Software-corrected molecules
with singletons removed.

K. Karlsson, Stanford University Medical School/E. Dewalt, Nature Research

Molecular barcodes can help researchers track molecules and avoid errors due to amplification and sequencing.

UMIs and wobble

Many companies sell oligonucleotides and barcoding kits, with a variety of nucleotide sequence lengths. Custom DNA and RNA oligos and primers are available, and companies use different types of purification and quality control techniques.

Eurofins Genomics, for example, sells synthetic oligos. The company is part of Eurofins, which does testing for the pharmaceutical, food and environmental industries. Sequencing technology has steadily moved into the company's testing activities, says Philipp Wenter, who heads Eurofins Genomics' research and development, such that his labs are users of high-throughput sequencing as well as troubleshooters for oligo synthesis methods. The company's research-based customers work in different types of high-throughput sequencing labs and in synthetic biology.

When labs decide about using oligos as molecular identifiers, they consider size, scope and type of experiment, and cost, says Wenter. Motivated to avoid PCR duplicates, many labs want to try out different label designs with smaller reagent amounts. To address this need, Wenter and his team tweaked synthesis chemistry and engineering. Typically, oligo companies will provide around 10 nanomoles of primer, he says. "We are able to produce about a fifth or a tenth of that amount routinely and with high-throughput synthesis at the same quality and turnaround time."

Wenter sees much activity in single-cell-oriented labs and notes a focus on oligo design and on keeping oligos as short as possible to save space for sequencing genomic information of interest, he says. Labs often assume oligos are barcodes with randomized sequences. "As everything in nature, including chemistry, nothing is perfect, nothing is 100%," he says. During oligo synthesis, a chemical spelling mistake can occur when a base is skipped in a fraction of the oligo population.

Oligo companies, including Eurofins Genomics, are concerned about bias presented by both chemistry-related nucleotide spelling mistakes and issues related to synthesis with 'wobble' or 'degenerate' bases, says Wenter. The latter refers to the fact that a random base at a desired location can be one of four bases. But each A, C, G and T phosphoramidite building block has its own kinetics, and the coupling speeds can vary with synthesis conditions. The oligo, the UMI-to-be, is therefore not entirely random.

Chemical spelling mistakes and wobble-related biases are both low-frequency events and cannot be addressed with the standard quality control measures that companies have used for decades. Neither mass spectrometry nor analytical high-performance liquid chromatography has sufficient sensitivity for such variations. "What we find more and more is that the actual quality control is the experiment itself," he says. The best test is the actual high-throughput-sequencing-oriented library prep, followed, in some cases, by sequencing.

As new techniques and experimental needs emerge, Wenter's sense is that manufacturers are watchful about possible bias sources, and troubleshoot them. He openly communicates with the research community also about possible product limitations, so he and his team can then keep tweaking the oligo-synthesis workflow.

Sensitivity focus

Experiments vary in terms of the desired sensitivity for mutation detection, says Illumina's Schroth. Labs performing amplicon sequencing and seeking 0.1% base-calling sensitivity need many reads and error-correction methods such as the use of UMIs to ensure base-calling confidence. To 'call' at 0.1%, one needs around 5,000× sequencing depth to be sure a mutant allele is seen a few times. An experimenter needs 5–10 molecules for each read to 'collapse' the data during analysis and eliminate all errors that might have occurred during PCR and sequencing. The Illumina spin-off company GRAIL is focused on highly sensitive assays that might be used in cancer screening to analyze cell-free DNA in blood samples. These assays "absolutely require UMIs; you can't do that assay any other way," says Schroth.

Dmitry Shcherbo

Dmitriy Chudakov hopes that soon all resequencing experiments will use unique molecular identifiers (UMIs).

With cell-free DNA analysis to, for example, explore minimal residual disease in cancer, a lab might use tools such as Safe-SeqS, developed in the Vogelstein lab at Johns Hopkins University, in which PCR duplicates are collapsed to form a consensus, which gives better results, says Schroth. In targeted cancer panels, in which labs look at certain cancer-related genes, sequencing depths might reach 30,000× or more. There are perhaps ten PCR duplicates for every molecule. The duplicates are combined computationally into a consensus read, which helps sort out errors. This approach, says Schroth, "tells you with really high confidence what that molecule was that has that particular UMI on it."

Most scientists use Q30 bases, says Schroth, which means that the probability of an incorrect base call during sequencing is around 1 in 1,000, providing base-calling accuracy of 99.9%. With Q30 bases, when single-nucleotide polymorphisms are called without UMIs, the average error rate is 0.1–0.2%, but the rate varies across individual sequences. For applications such as looking for resistant clones in a cancer patient, researchers will want to eliminate sequencing noise at levels that are 100-fold higher, which is a level that UMIs can deliver, he says.

Single-cell analysis lets labs study a sample's cellular heterogeneity in ways not possible with bulk transcriptome analysis. With such experiments in mind, Illumina and Bio-Rad developed and launched a system that uses UMIs to count unique mRNA molecules, says Schroth. Researchers can do multiplexed experiments to look at several samples under varying conditions and at different time points. The process yields more than just the typical RNA-seq result of reads that map to the transcripts. It delivers a more quantitative result, it gives transcripts per cell, he says.

In the workflow, single cells are separated and partitioned into tiny droplets, the mRNA molecules from each cell are barcoded with UMIs, the droplets are burst and the cDNA is pooled. Next comes cDNA tagmentation and library prep followed by sequencing. There is no DNA shearing; instead, there is one-step enzymatic fragmentation and the attachment of sequencing adaptors. A proprietary engineered transposase holds on to the oligos carrying the sequencing adaptors, and when the transposase binds to a cleaved genomic fragment, it attaches the adaptors and skips the typical adaptor-ligation process.

Standard RNA-seq requires around 100 nanograms of RNA, which is sometimes more than a lab has. At Illumina, Schroth and his team have found that labs can work with much less; even with 10–20 nanograms, the library will be of good quality. But labs will have to deal with the higher number of PCR duplicates these experiments will involve compared to those with more starting material.

Eurofins

Philipp Wenter and his team have tweaked oligo synthesis to better deliver small reagent amounts.

In the Illumina labs, the team also experimented with purposefully generating a large number of PCR duplicates. The team compared data from unique reads and duplicates—'good' and 'bad' data. "Essentially—I was even a little surprised by this—you couldn't really tell the difference; the good and the bad data were identical," says Schroth. This experimental outcome reinforced his notion that under certain conditions, such as typical RNA-seq assays, PCR duplicates are not problematic. But this is not the kind of assay that might be performed in cancer diagnosis or treatment.

Speaking generally on the subject of sequencing-library prep and PCR, Schroth says that "people are way more paranoid about it, for the most part, than they need to be." Thwarting PCR duplicates in single-cell analysis is, however, "a completely valid topic," he says. Dealing with PCR duplicates is essential for highly sensitive assays with low levels of input material, and "a single cell is the ultimate low input in the RNA world."

Vivien Marx is technology editor for *Nature Methods* (v.marx@us.nature.com).

1. Karlsson, K. *Counting Molecules in Cell-free DNA and Single Cells RNA*. PhD thesis, Karolinska Institutet (2016).
2. Kivioja, T. *et al. Nat. Methods* **9**, 72–74 (2011).
3. Hug, H. & Schuler, R. *J. Theor. Biol.* **221**, 615–624 (2003).
4. Miner, B.E., Stöger, R.J., Burden, A.F., Laird, C.D. & Hansen, R.S. *Nucleic Acids Res.* **32**, e135 (2004).
5. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W. & Vogelstein, B. *Proc. Natl. Acad. Sci. USA* **108**, 9530–9535 (2011).
6. Ebbert, M.T. *et al. BMC Bioinformatics* **17**, 239 (2016).
7. Shugay, M. *et al. Nat. Methods* **11**, 653–655 (2014).
8. Smith, T., Heger, A. & Sudbery, I. *Genome Res.* **27**, 491–499 (2017).