Understanding and improving high-throughput sequencing data production and analysis

Martin Kircher

Summary

New high-throughput sequencing technologies make it possible to apply sequence-based approaches in an unanticipated number of fields. In the field of evolutionary genetics, it is now feasible to apply sequencing-based approaches for a wide range of comparative genomic studies. For example, high-throughput sequencing can be applied to study the genomes from ancient specimens of different hominin groups, like Neandertals and Denisovans, and allow large-scale population genetics studies of present-day humans as well as measuring quantitative differences in the transcriptomes and DNA-interactome of different apes and primates.

However, while the cost and time for applying these new technologies was greatly reduced compared to traditional Sanger sequencing, the error profiles and limitations of the new instruments differ significantly from those of previous technologies. Further, the types of errors observed as well as the number and length of sequences vary considerably between these different new technologies. Therefore, data analysis requires a detailed understanding of the imperfections in the resulting sequence data and how these pose challenges and cause biases. I review current sequencing technologies and point out their conceptual limitations. In this thesis I describe very specific biases and limitations, which go back to the technical details of how DNA molecules are prepared for sequencing, sequencing templates immobilized and finally read out.

Current high-throughput technologies have an average error rate of 1/25 to 1/1,000, which is considerably higher than the 1/10,000 to 1/100,000 observed for high quality Sanger sequence read outs. The *in vitro* amplifications which are generally performed prior to sequencing introduce a higher error into the sample before it enters the actual sequencing process. In addition, currently used random-dispersal protocols for immobilization of sequencing templates using beads or other solid surfaces cause mixed signal read outs and dependence of sequencing errors from strength and distance of close-by sequencing reactions.

Most errors on the new instruments originate from signal-to-noise thresholding and signal detection issues. Further, error rate substantially increases with the position in the sequence due to reductions in reaction efficiency, molecule damage and phasing, a process in which not all molecule copies are equally extended in every sequencing step. Shorter read lengths from these new platforms limit the accurate sequence mapping and assembly of genomes. Only paired end or mate pair protocols help to overcome some of these limitations by providing information about relative location and orientation of a pair of reads.

I have analyzed the currently most frequently used high-throughput sequencing platform, the Illumina Genome Analyzer, in more detail. Based on the problems observed frequently in runs performed at the Max Planck Institute for Evolutionary Anthropology, I present simple rules, which shall enable the identification and handling of the most common problems. I describe the different sources of high variance in run quality, ranging from issues with the sequencing libraries, to incorrect instrument adjustment and handling. Particles like chemistry lumps, dust and lint can cause pseudo sequence signals which result in the analysis of low sequence complexity reads not originating from the actual sequencing library. While sequence entropy filters efficiently remove these sequence, tagging or indexing allow a superior method for filtering real library molecules and further reduce the risk of library contamination.

For sequence reads where part of the adapter sequence is included, the position in the se-

quence read at which the adapter sequence begins has to be identified and the read trimmed appropriately. Unfortunately, this is not part of the standard Illumina data processing and also non-trivial for short adapter fragments, especially given the increasing sequencing error at the end of reads. If reads are not filtered for known chimeras and trimmed for adapter sequences, these may interfere with mapping/alignment and thereby impact downstream analysis. For paired end reads the correct identification of the adapter set-in is eased by maximizing autocorrelation of the two reads with the outlined read merging process. In addition to the efficient identification of adapters, merging reduces error rates in the consensus called sequence part. The algorithm presented has therefore been vital for different ancient DNA studies at the Max Planck Institute. Various library preparation biases may exist and impact sequencing results. For this reason, for example PCR duplicates need to be identified and specifically handled in analysis.

Considering that differences in error profiles are one of the major differences between technologies, reduction of these errors and precise estimates for the correctness of a specific base in a sequence are very important for any type of analysis. I present a new approach to base calling, the conversion of intensity measures into bases, for Illumina sequencing instruments. The approach presented is unique and currently applies to the full range of different Illumina sequencing chemistries and platform versions, for which it reduces sequencing error by at least 10-20%.

On the Illumina platform a strong correlation of adenine and cytosine intensities and of guanine and thymine intensities as well as a dependence of the signal for a specific cycle on the signal of the cycles before and after (phasing and pre-phasing) complicate base calling. Previous approaches have either completely modeled the sequencing process or at least corrected raw intensities prior to the application of statistical learners. Therefore, all these approaches depend on a good understanding and modeling of the sequencing process. The developed base calling package, `Ibis` (`Improved base identification system`), by-passes this problem by direct training of one statistical model per sequencing cycle based on raw cluster intensities of multiple input cycles, directly incorporating the effects of phasing. Thus, `Ibis` implements the most general and flexible approach, which is of advantage when considering the vast improvements of sequencing chemistry and instrument over the last years. Further, the performance of `Ibis` on standard hardware is significantly better than for other existing alternative base callers. Increases in mappable sequences due to reduced base identification errors as well as improved and calibrated `PHRED`-like quality scores enable the direct use of the sequences in other software packages.

I present two applications of `Ibis` and other principals presented in this thesis. I analyzed one of the first applications of the Illumina sequencing platform, the *NlaIII* Digital Gene Expression (DGE) approach, which infers gene expression levels through short 17nt-tag sequencing of the 3' ends of transcripts. This protocol was used to study brain, heart, kidney, liver and testis tissues of humans, chimpanzees and rhesus macaques. The biggest analysis challenge were the short tags which are not unique to specific genomic sites or genes and for which the uniqueness of tags differs slightly between the three species. Further, annotation of tags was problematic due to very different annotation quality for the three species. Only very recent human gene annotation provided the necessary annotation of 3' untranslated regions and could be projected to the chimpanzee and rhesus macaque genomes, losing about 36% of genes annotated in human but giving similar proportions of tag counts within genes for all three species.

From comparisons to other studies of the same species and tissues, larger disagreement was observed than was expected. For example, differences in the percentage differentially expressed genes or in the symmetry of assignment of changes to evolutionary lineages were observed. It is likely that all methods have technological (experimental and analysis) biases. A comparison with the Babbitt et al. study, also using the *NlaIII* DGE protocol but for different brain samples, clearly shows that sampling variation is at least in the range of biological differences between human and chimpanzee and that analysis variation may even be as strong as differences of humans and chimpanzees when compared to rhesus macaques. Future studies will need to control sample environmental effects, sample age, and tissue sampling more thoroughly. Further, improved experimental and analysis protocols are required which allow to detect and measure subtle effects that could introduce a species bias. Currently, species specific differences may easily originate from different genome quality, genome completeness and genome annotation quality.

The second analysis presents whole genome shotgun sequencing data that was generated for two hominin genomes from ancient DNA, the Neandertal and Denisova genome. Ancient DNA sequences are generally short in length, damaged, and at low copy-number relative to co-extracted environmental DNA. For Neandertal and Denisova the challenges from sequencing ancient DNA, which include adapter sequence at the read ends, chimerical sequences and other artifacts as well as sequencing error for short molecule lengths, have been addressed using the described approaches of improved base calling, tag filtering, and short paired end read merging. In combination with experimental approaches for reducing ancient DNA damage and the consensus from PCR duplicates, the sequencing error associated with ancient DNA studies could be considerably reduced. The remaining error from sequencing and error originating from ancient DNA damage in the Denisova molecule read outs is even lower than for present-day human sequences generated with the same technology.

I show how the ancient DNA sequences can be used to study sites in the human genome which have changed since the last common ancestor of human and chimpanzee and to identify features that set fully anatomically modern humans apart from other hominin forms. The identified positions point to several regions and genes, some of which might be affected by positive selection in the recent evolutionary history of modern humans. Experimental work will be required to elucidate the physiological consequences of the identified changes. In addition, I describe an interesting subset of sites which changed on the human lineage. I identified tens of thousand of positions where Denisovans and Neandertals disagree in the ancestral state at sites where the human reference sequence carries the derived allele. These positions are inconsistent between lines that separated more than half a million years ago and at least partially, reflect variation at the point of Human-Neandertal-Denisova lineage separation that segregated differently in the three lineages (incomplete lineage sorting).

It is likely that a large proportion of these sites, which were polymorphic at the time when human, Neandertal and Denisovan lineage separated, fixed for the derived allele in present-day humans. Thus, these differently segregating sites might have been reintroduced into some present-day human populations by admixture with either Neandertals or Denisovans and can be used to test present-day human individuals whether they show more frequently the ancestral allele for the Denisova ancestral sites or the ancestral allele for Neandertal ancestral sites. When analyzing these Neandertal-Denisova discordant sites in twelve present-day populations, they turned out to be informative for detecting admixture with either of the ancient population. I could confirm that an African individual shares fewer ancestral alleles with Neandertal than do all non-African individuals, supporting the admixture signal with non-Africans described in Green et al. for the Neandertal genome. Further, I could show that Melanesians, especially the two Papuan individuals, show a signal of Denisovan admixture not shared with other sampled populations, a result in agreement with the D-statistics for population pairs presented in Reich et al. for the Denisova genome.

During analysis, I point out differences in sampling of the reads obtained for the Neandertal and the Denisovan genomes. For example, in human accelerated regions Neandertal and Denisova data both show that these regions tend to predate the human-Denisova-Neandertal split and that differences caused by biased gene conversion tend to be older in time, however we sampled much more reads covering ancestral sites just in the Denisova data. While this may point to a simple sampling effect, an excess in the number of Denisova ancestral sites was observed in the concordance analysis, which can not result from sampling and also does not originate from a human reference sequence bias (from the Neandertal admixture present in parts of the reference genome). Currently this excess might either orignate from different alignment approaches or admixture into the Denisovan individual from some archaic hominin.

Both analyses pointed out that small effects throughout the whole data generation and data analysis process may introduce sufficiently large biases to complicate drawing biological conclusions from experimental data. The information and approaches outlined in this thesis, will however help to either prevent generating such biased data sets or at least reduce the sequencing instrumentation biases.