A caution to industry and regulators - "Incidental DNA fragments" may be misinterpreted using Next Generation Sequencing (NGS)

S. Newmaster, S. Ragupathy and R. Hanner

Dr Steven Newmaster
Professor, Botany & Genetics/Genomics
University of Guelph, Canada
snewmast@uoguelph.ca

Dr Subramanyam Ragupathy
Senior Scientist, NHP Molecular Diagnostics
Biodiversity Institute of Ontario, Canada
ragu@uoguelph.ca

Dr Robert Hanner
Chief Technology Officer (CTO), TRU-ID Ltd
Research Park Centre, Suite 102, 150 Research Lane, Guelph, Ontario N1G 4T2
rhanner@tru-id.ca
http://www.tru-id.ca

DNA is everywhere. Unless you are reading this in a completely sterile room, you are surrounded by small fragments of the nucleotides that make up your own DNA and that of all of your friends who have recently visited, as well as the DNA of your pets, each of the plant and animal ingredients in your lunch, and maybe even your clothing if you are wearing cotton, wool, or other natural fibers. Trace amounts of DNA of all of the plants in your vicinity may also be scattered around you if you are in a location that windblown pollen can reach.

Such "incidental DNA fragments" should also be assumed to be present in various combinations and amounts in samples of herbal ingredients. In fact, pharmacopoeial botanical monographs and industry standards specifically allow some small amount – usually 2 to 5 percent – of "foreign organic matter," [1,2] which may reasonably include other plant parts of the target species or inadvertent but minimal presence of other species that may be co-mingled in a harvested crop. Incidental DNA can also be found from a trace of another plant's DNA on the surface of the leaves, flowers, roots and stems of the any herb intended for use as in ingredient in a manufactured finished product.

Testing of plant species and ingredients derived from plants to verify identity is always a challenge, and certain criteria must be met irrespective of the particular tools used. Whether plant identity is determined by sensory or organoleptic characteristics, by observation of distinctive characteristic macroscopic or microscopic features, by chemical analysis, or with DNA techniques, the specific test(s) must be performed by individuals qualified by training and experience; must be "fit-for-purpose" to ensure each test is relevant to the specific tested material; must rely on authentic references; and must be scientifically valid.

There are particular and unique challenges associated with using DNA tools for verifying herbal ingredient identity, and we have some concerns that we will share in this article involving incidental DNA and one type of DNA analysis known as next generation sequencing (NGS).

NGS is a powerful research tool, as it is very useful for detecting multiple sources of DNA in a single analysis, a quality that can be useful in answering certain research questions. For example, this technique has allowed a whole new discipline of research into the study of traces of environmental DNA (eDNA), that exists in water and soils[3,4]. This has applications, for example, in monitoring for invasive fauna species in waterways[5].

But the published literature indicates there are considerable problems with NGS[6,7] that present an immediate impediment to generating scientifically valid test results, as are necessary for commercial use of this tool to verify herbal ingredient identity, and that therefore require additional research. Notably, NGS results may indicate presence of species in a sample due only to detection of incidental DNA, and may also over-estimate the amounts of incidental DNA. Such reports result in conclusions that are irrelevant at best and potentially misleading, and in fact may only point to a trace amount of an inconsequential taxon that has no real relevance to the analysis or underlying research[8].

There is considerable additional primary research literature suggesting several other challenges that must be overcome to make NGS most useful and accurate, including polymerase chain reaction (PCR) amplification bias[8], which may skew estimates of species abundance. Other general NGS issues that need to be addressed through additional research include low-coverage, repetitive sequence, pseudogenes, homopolymer repeats, and large insertions and deletions[6,7]. Such questions concerning the limitations of NGS have yet to be fully addressed despite considerable effort by very large research institutes.

One solution to some of these challenges would be to develop bioinformatic algorithms that adjust the estimates of sequence abundance from NGS per species. This includes current research on targeted NGS, which can be successful in closed systems with a small number of known species[10,11]. However, in industries where contaminants and adulterants are not well-known, and where hundreds of herbal species are used, this will be a monumental task[12]. Moreover, NGS-based sequence matching and discovery is very challenging without well-developed DNA reference libraries that include samples of closely related non-target species, all with good populations sampling[13]. To date, scant research has been published on NGS standard operating protocols (SOPs), libraries and pipelines for the analysis of dietary supplements as marketed in the U.S., or Natural Health Products (NHPs) as sold in Canada. Rather, the current research priorities for NGS are focusing on detection of human and food pathogens and these tools are expected to be commercially available within a few years[14].

It is useful to consider as an aside why the foodborne pathogen industry has not yet adopted NGS technology[15,16]. In short, they have not yet been able to overcome the above issues in order to provide a statistically valid test[7,8,17]; and when applied to detection of food pathogens the possibility of a false positive result can cause significant business

disruption, while any chance of a false negative is well understood to be an unacceptable threat to public health. Progress in this area is being made, however, and current research suggests commercial NGS tests for foodborne pathogens will be available within the next two years[15].

NGS is a promising technology that will be commercially available for the food, supplement, and NHP industries once the libraries and pipelines have been developed, validated, and published in peer-reviewed scientific journals[16]. Our research team is contributing to the necessary research and development initiatives for NHPs, but community accepted methodological validation of such methods is still likely to be several years away. Even when completely validated, the use of NGS approaches will still require relatively expensive equipment operated by highly qualified personnel in ISO accredited test labs, or perhaps on-site by manufacturers with sufficient QA/QC budgets[8,18].

Research presented at the recent AOAC annual meeting[19] cautions against the use of NGS as this technology may lead to false positives for adulterants and contaminants in NHPs. The presentation shared results of experimental mixtures of botanical ingredients (herbarium vouchers) that were generated using NGS and analysed in several internationally recognized labs. The results indicated that NGS often missed species and/or added species to the list of NGS test results, which did not match the experimental mix of botanical ingredients. Furthermore, the quantitative estimates from NGS mixtures did not match any of the experimental mixtures.

During the past year we have responded to industry concerns that commercial NGS test results are reporting confusing results including considerable weed species in samples. This is not surprizing since farm operations encounter considerable amounts of incidental DNA from agricultural weeds in every field; NGS is so sensitive it is possible it is detecting very small numbers of DNA fragments from weeds on a farm that are in only negligible amounts in the harvested crop – so again, the presence of incidental DNA appears to be confounding the accuracy, or at least the marketplace relevance of these tests. We are conducting research on this phenomenon and will be publishing on it soon. Furthermore, some commercial NGS test results indicate several adulterants present in a sample that was known to be a simple, single, identifiable leaf. NGS detects incidental DNA fragments commonly present in manufacturing facilities and this analytical noise also needs to be controlled for when using NGS, as the tiny quantities likely present do not impact the quality of a dietary supplement or an NHP product. Another explanation for confusing and incorrectly interpreted NGS results is mismatching unknown DNA sequences to poorly developed libraries[13] or even the use of genus specific primers that are known to incorrectly match to a number of closely related species; this is more evidence of poorly conceived NGS pipelines[7,12].

Although incidental DNA is a considerable problem for NGS techniques, there are commercial DNA-based identification tools that have been tested and validated for commercial use. These tools have been adopted by the food borne pathogen industry[20,21,22] and have been developed for commercial plant species used in food[16] and in dietary supplements and NHPs. This work was based originally on a DNA barcoding initiative[23] led

by the University of Guelph in which we have developed SOPs[24] and extensive DNA libraries for many species and populations, including NHPs[25] – which have known provenance and sample vouchers stored in our collections facility. To date we have conducted over 100,000 DNA-based tests that have served in the development of rigorous statistical models and the development of an extensive Biological Reference Material Library (BRM) that can be accessed by industry partners, and regulators. This initiative supports the call for pharmacovigilance of NHPs[26] and addresses concerns in our early research[27] by providing novel molecular diagnostic tools that we have been beta-testing with leading industry partners. Our vision is to develop an alliance with industry leaders, regulators and consumers in the development of new industry testing standards to serve those who seek quality ingredient supply chains as we focus on research and development of reliable, affordable DNA-based tools for validating species ingredients.

1. United States Pharmacopeia—National Formulary (USP—NF). (2015). Astragalus Root Powder. The United States Pharmacopeial Convention, Rockville, MD, USA.
2. American Herbal Products Association. October 2017. Guidance on Limits of Foreign Matter in Herbal Ingredients. AHPA: Silver Spring, MD.
3. Carew, M.E., Pettigrove, V.J., Metzeling, L., and Hoffmann, A.A. (2013). Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species. *Frontiers in zoology*, *10*(1): 1.
4. Hajibabaei, M., Spall, J.L., Shokralla, S., and van Konynenburg, S. (2012). Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. *BMC ecology*, *12*(1): 1.
5. Rees, H.C., Maddison, B.C., Middleditch, D.J., Patmore, J.R.M, and Gough, J.C. (2014). The detection of aquatic animal species using environmental DNA – a review of eDNA as a survey tool in ecology. *J Applied Ecology*, *51*:1450-1459.
6. Coissac, E., Riaz, T., and Puillandre, N. (2012). Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular ecology*, *21*(8): 1834-1847.
7. Schrijver, I,, Aziz, N., Farkas, D.H., Furtado, M., Gonzalez, A.F., Greiner, T.C. and et al. (2012). Opportunities and challenges associated with clinical diagnostic genome sequencing: a report of the Association for Molecular Pathology. *J Mol Diagn*, *14*(6): 525–40.
8. Burns, M., Wiseman, G., Knight, A., Bramley, P., Foster, L., Rollinson, S., Damant, A., and Primrose, S. (2016). Measurement issues associated with quantitative molecular biology analysis of complex food matrices for the detection of food fraud. *Analyst*, *141*: 45-61
9. Thudi, M., Li, Y., Jackson, S.A., May, G.D., and Varshney, R.K. (2012). Current state-of-art of sequencing technologies for plant genomics research. *Briefings in Functional Genomics*, *11*(1): 3-11.
10. Metzker, M.L. (2010). Sequencing technologies—the next generation. *Nature Reviews Genetics*, *11*(1): 31-46.
11. Stefan, C.P., Koehler, J.W., and Minogue, T.D. (2016). Targeted next-generation sequencing for the detection of ciprofloxacin resistance markers using molecular inversion probes. *Scientific Reports*, *6:25904*:1-12.

12. Claros, M.G., Bautista, R., Guerrero-Fernández, D., Benzerki, H., Seoane, P., and Fernández-Pozo, N. (2012). Why assembling plant genome sequences is so challenging. *Biology*, *1*(2): 439-459.
13. Egan, A.N., Schlueter, J., and Spooner, D.M. (2012). Applications of next-generation sequencing in plant biology. *Am J Botany*, *99*(2): 175-185.
14. Ripp, F., Krombholz, F., Liu, Y., Weber, M., Schafer, A., Schmidt, B., Koppel, R., Hanklen, T. (2014). All-Food-Seq (AFS): a quantifiable screen for species in biological samples by deep DNA sequencing. *BMC Genomics*, *15*:639.
15. Corrado, G. (2016). Advances in DNA typing in the agro-food supply chain. *Trends in Food Science & Technology*, *52*: 80-89.
16. Cavin, C., Cottenet, G., Blancpain, C., Bessaire, T., Frank, N., and Zbinden, P. (2016). Food Adulteration: From Vulnerability Assessment to New Analytical Solutions. *CHIMIA International Journal for Chemistry*, *70*(5): 329-333.
17. Downey, G. (2016). Advances in food authenticity testing. *Woodhead Publishing series in food science, technology and nutrition*.
18. Mishra, P., Kumar, A., Nagireddy, A., Mani, D.N., Shukla, A.K., Tiwari, R., and Sundaresan, V. (2016). DNA barcoding: an efficient tool to overcome authentication challenges in the herbal market. *Plant Biotechnology J*, *14*(1): 8-21.
19. Newmaster, S.G. (2016). *Botanical species ingredient validation using DNA-based molecular diagnostic tools*. In, Botanical Dietary Supplement Ingredient Identity Authentication: Determining Appropriate Reference Materials for Botanicals and Their Extracts. AOAC 130th Annual Meeting & Exposition, Dallas, TX, September 18-21, 2016.
20. Cammà, C. et al. (2012). Development and validation of fast Real-Time PCR assays for species identification in raw and cooked meat mixtures. *Food Control*, *23*: 400-404.
21. Malorny, B. et al. (2004). Diagnostic Real-Time PCR for Detection of Salmonella in Food. *Appl Env Microbiol*, *70*: 7046–52.
22. Heller, L.C. et al. (2003). Comparison of Methods for DNA Isolation from Food Samples for Detection of Shiga Toxin-Producing Escherichia coli by real-time PCR. *Appl Env Microbiol*, *69*(3): 1844–46.
23. Newmaster, S.G., Fazekas, A., Ragupathy, S. (2006). DNA barcoding in the land plants: evaluation of rbcL in a multigene tiered approach. *Canadian J Botany*, *84*:335-341.
24. Fazekas A.J., Kuzmina M., Newmaster S.G., and Hollingsworth P.M. (2012). DNA Barcoding Methods for Land Plants [invited chapter]. In: DNA Barcodes: Methods and Protocols. Edited by WJ Kress and DL Erickson. Springer, New York, USA.
25. Newmaster, S.G. and Ragupathy, S. (2014). Ethnobotany Genomics. [invited chapter]. In: Recent Advances in Plant-Based, Traditional and Natural Medicines. Edited by S. Vemulpad and J. Jamie. CRC Press, Taylor & Francis, Waretown, NJ, USA.
26. de Boer, H.J., Ichim, M.C., and Newmaster, S.G. (2015). DNA barcoding and pharmacovigilance of herbal medicines. *Drug Safety, 1-10 DOI 10.1007/s40264-015-0306-8.*
27. Newmaster, S.G., Grguric, M, Shanmughanandhan, D., Ramalingam, S. and Ragupathy, S. (2013). DNA barcoding detects contamination and substitution in North American herbal products. *BMC Medicine*, *11*:222-235.