# EXHIBIT A

UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

| | |
|---|---|
| CHEVRON CORPORATION,<br><br>Plaintiff,<br><br>-against-<br><br>STEVEN R. DONZIGER, et al.,<br><br>Defendants. | Case No. 11 Civ. 0691 (LAK) |

**DIRECT TESTIMONY OF PATRICK JUOLA, Ph.D.**

I, PATRICK JUOLA, hereby declare under penalty of perjury pursuant to 28 U.S.C. § 1746, that the following is true and correct:

1. I am a tenured Associate Professor of Computer Science at Duquesne University, in Pittsburgh, PA. I am an expert in computational and forensic analysis, specifically related to text and authorship analysis. I have authored 40 peer-reviewed articles, primarily on the computational inference of document authorship via the statistical analysis of linguistic features. In my research, I focus on the computational and forensic analysis of linguistic features, and I specialize in the area of authorship attribution.

2. I have been retained by Gibson, Dunn & Crutcher, LLP ("Gibson Dunn") on behalf of Chevron Corporation ("Chevron") in this case to determine whether certain documents of the lawyers and consultants for the plaintiffs in the case of *Maria Aguinda y otros v. Chevron Corporation* (also known as the Lago Agrio case) can be found in the trial court record.

PLAINTIFF'S EXHIBIT
3800
11 Civ. 0691 (LAK)

## Summary of Expert Opinion

3. Based on my expert computer-based textual analysis of the Lago Agrio plaintiffs' lawyers and consultants' work product identified in the Ecuadorian judgment and the trial court record in the Lago Agrio case, I have concluded, to a reasonable degree of certainty, that the Fusion Memo, the Clapp Report, the Index Summaries, the Fajardo Trust email, the Draft Alegato, and the Selva Viva Data Compilation are not in the trial court record. Moreover, I have concluded, to a reasonable degree of certainty, based on additional forensic evaluation, that the computer-based analysis was effective to identify sources for the linguistic overlaps between the plaintiffs' work product documents and the Ecuadorian judgment, and that those sources do not exist in the trial court record.

## Background and Qualifications

4. I am a tenured Associate Professor of Computer Science at Duquesne University, Pittsburgh, PA. I am also the Director of the Evaluating Variations in Language Laboratory, also at Duquesne.

5. I am also the founder and Director of Research for J Computing, Inc., (dba Juola & Associates "J&A"), a Pennsylvania corporation specializing in text and authorship analysis.

6. I obtained a Bachelor of Science degree in 1987 in electrical engineering at the Johns Hopkins University, Baltimore, where I double majored in Mathematics and Electrical Engineering. I also obtained an M.S. degree in Computer Science from the University of Colorado in 1991 and an M.S. level certificate in cognitive science, also from the University of Colorado, in 1993. Finally, I received a Ph.D. in computer science from the University of Colorado in 1995. I was a postdoctoral research associate for the Department of Experimental Psychology at St. Hugh's and Lincoln Colleges, Oxford University, from 1995 to 1998.

7. In 1998, I started at Duquesne, where I took the position of Assistant Professor of Mathematics and Computer Science. In 2004, I was granted tenure and became an Associate Professor. At Duquesne, I teach classes, including classes in natural language processing, logic programming, software engineering, and cryptography. In my research, I focus on the computational and forensic analysis of linguistic features. Within that, I specialize in the area of authorship attribution.

8. I was inducted into the Office of Research Hall of Fame at Duquesne in 2009. Also in 2009, I received the McAnulty College Faculty Excellence in Scholarship Award from Duquesne.

9. I have authored over 150 publications in general, 40 of which have been peer-reviewed articles. I've also written 2 books and 9 book chapters. Most of my publications discuss the computational inference of document authorship via the statistical analysis of linguistic features.

10. I am a frequent ad-hoc reviewer on subjects pertaining to authorship attribution, stylometry, digital humanities, and text analysis for a number of journals, including LLC (formerly Literary and Linguistic Computing), JASIST (Journal of the American Society for Information Systems Technology), and SPE (Software Practices and Experiments).

11. My company, Juola & Associates, specializes in computational and forensic analysis, specifically with text and authorship analysis. We handle projects involving authorship attribution, profiling of author characteristics, plagiarism, and also large scale document searching.

12. I am the primary architect and designer of the JGAAP (Java Graphical Authorship Attribution Program) authorship analysis system. This system, funded by the National Science

Foundation (NSF) for nearly $2 million, is a system for developing and testing new methods of authorship attribution and determining best practices. In addition to developing best practices, the NSF has also charged me with the development of a forensic-quality authorship attribution system (a goal we have met with the current version of JGAAP) as well as commercializing the technology developed for this purpose.

**Methodology**

13. I first reviewed the findings of other experts in this case, including those of Dr. Robert Leonard, who had identified linguistic overlaps between the plaintiffs' work product documents and the Ecuadorian judgment. Dr. Leonard identified at least nine instances of substantial textual overlap between the Ecuadorian judgment and plaintiffs' work product documents.[1]

14. Dr. Leonard identified linguistic overlap between the Ecuadorian judgment and the following plaintiffs' work product documents, which I was provided by counsel for Chevron:

a. A document entitled "Primer Borrador Memo Fusión JSP [Nov2007].doc" (henceforth the "Fusion Memo," PX 435);

b. Two versions of an unfiled spreadsheet entitled "pruebas pedidas en etapa de prueba.xls" and "GARR-HDD-003243" (henceforth the "January Index Summary," PX 433, and the "June Index Summary," PX 434, and collectively the "Index Summaries");

c. A draft trial brief, known as an *alegato*, containing the allegations and arguments of the Lago Agrio Plaintiffs (henceforth the "Draft Alegato," PX 438);

[1] *See* June 27, 2011 Report of Dr. Robert Leonard, pg. 11.

d. An email from "Pablo Fajardo Mendoza" to three people including "Steven Donziger" on the subject of "FIDEOCOMISO" (henceforth the "Fajardo Trust email," PX 437);

e. An email (DONZ00025295.pdf) forwarding a report containing the text of DONZ00025296.doc (henceforth the "Clapp Report," PX 928);

f. A document containing sampling data known as the "Selva Viva Data Compilation" (PX 439-41).

15. I received the trial court record in two stages. In the first stage, on or about September 13, 2011, Chevron's counsel provided J&A with approximately 3500 electronic documents comprised of approximately 236,000 images of individual pages, numbered CL0000-00000 through CL2068-0216692, with the February 14, 2011 Judgment beginning at page CL2065-0216338. I understood from Chevron's counsel that this version of the record I reviewed is from a photocopy of the official version of the record maintained by the Provincial Court of Justice of Sucumbios, and that this photocopy was prepared by the Clerk of Court of the Provincial Court of Justice of Sucumbios per normal court procedures, stamped with a court seal on each page to indicate authenticity of the copy, and delivered to Chevron in installments as requested by the company's Ecuadorian trial counsel over the course of the lower court trial. Chevron scanned these copies, creating PDFs, which were then converted to single-page TIFF format and uploaded to an electronic platform, at which point the files were subjected to an automatic OCR process.

16. On or about May 30, 2013, J&A received from Chevron a hard drive containing the electronic contents of 69 compact discs (henceforth the "CD Content List") provided by the National Court of Justice of Ecuador. I understood from Chevron's counsel that the Attorney

General of Ecuador had previously requested copies of all digital information contained on CDs or DVDs in the court record and that Chevron had subsequently requested its own copies of this digital information.

17. Due to the size and heterogeneity of the data received, our first task was to systematically convert all documents to a common format called UTF-16. This format is a variation of "plain text" but that allows for non-English letters or letters with diacritical marks (accents such as "ó"). This provides a basis to search based on words or characters using a common encoding in machine-readable form. In the process of this conversion, we also stripped out all punctuation and capitalization distinctions to maximize the chance of detecting matches between text identified in the plaintiffs' work product documents and the trial court record. This is a conservative procedure in that it ensures that words that differ only in capitalization or punctuation will be correctly matched.[2]

18. We broke each of the documents in the court record into word groups of length 5 (henceforth "5-grams" or more generally "n-grams"). In layman's terms, these are simply groups of five consecutive words. For example, the English phrase "Chief Justice of the Supreme Court of the United States" constitutes a 10-gram in its own right and contains within it many overlapping 5-grams, including "Chief Justice of the Supreme," "of the Supreme Court of," and "Court of the United States."

19. To account for the possibility of OCR errors, we also created another set of "fuzzy" n-grams that treated all non-Latin characters (including characters with diacritical/accent marks) as alike, so the character "ó" would be considered to be identical to the character "ò" or

---

[2] The Ecuadorian judgment document itself is of course part of the court record, but comparisons of the Ecuadorian judgment with itself would not have been useful for finding the sources from which it derived, and hence it was removed prior to analysis.

for that matter, the character "ö." This treatment is done to help compensate for potential errors introduced by the OCR process. Accents are among the most fragile aspects of writing when subjected to OCR, as a bit of stray dirt on the lens or a bad printer/copier can easily introduce, change, or eliminate stray marks that will be interpreted as accent marks. By treating characters that differ only in diacritical marks as being the same, the effect of such errors on the analysis is greatly reduced or minimized.

20. N-grams are highly individual; it is uncommon to see matches of 7-grams or longer except in the cases where the n-grams are part of a common overlapping phrasal vocabulary. "Chief Justice of the Supreme Court of the United States" is an example of such a phrase, familiar to any lawyer. "President de la Corte Superior de Justicia de Nueva Loja" is another example, perhaps equally familiar to an Ecuadorian lawyer. Direct and attributed quotation, of course, would be another valid reason for two documents to share n-grams.

21. We then compiled a list of every specific linguistic overlap between the plaintiffs' work product documents and the Ecuadorian judgment identified by Dr. Leonard (henceforth "Examples"). We broke the Examples down into 5-grams as well.

22. Once documents were broken down into n-grams of five words (5-grams), we used computer software to identify any 5-grams that were shared between the Examples and the court record.

23. Based on these comparisons, we were able to find any documents in the court record that contained an exact match (without regard to diacritical marks) of at least five words with one of the Examples. For each of these documents, we also were able to identify an area of maximal similarity, describing the approximate degree of overlap and allowing us to look at the

specific instances to determine whether the result indicated a source document for the overlapping text.

24. If the computer identified any such matches, we first verified the match by visually comparing the matching phrase and the corresponding part of the court record. We then checked whether the match was a direct quotation. Finally, we analyzed the match to determine whether it was a common or stereotyped phrase, judging partially on the phrase's frequency and distribution across documents and partially on our understanding of the phrase's meaning.

25. As an illustrative example, we consider the similarity cited as Example 1 in the June 27, 2011 report of Dr. Leonard.[3] Example 1 is a block of text with more than 90 identical words appearing in the Fusion Memo and in the Ecuadorian judgment. Upon comparison of this 90-word overlap with the court record, we found exactly eleven matches of five words or more across the entire three thousand plus documents in the court record. None of these were substantial; in fact, all eleven were exactly five words long, and eight of the matches were of the same five-word phrase "en el Ecuador como una," a common phrase that appeared in numerous documents and contexts. Based on this review, I conclude with a reasonable degree of scientific certainty that there is no document in the trial court record that is a possible source for the 90-word passage identified by Dr. Leonard as appearing in both the Fusion Memo and the Ecuadorian judgment.

26. As another example, we consider the similarity cited as Example 2 in Dr. Leonard's report. Example 2 is a block of text with an approximately 150-word overlap between the Fusion Memo and the Ecuadorian judgment. We found one example of a ten word overlap between the text in Example 2 and a document in the record identified as number CL0063-

[3] June 27, 2011 Report of Dr. Robert Leonard, pg. 11.

0006644.txt. The overlap was as follows: "bombas sumergibles en cinco pozos en el Campo Lago Agrio." The context in which the overlap appears is entirely different than in Example 2. Therefore, based on our review, neither CL0063-0006644.txt nor any other document in the trial court record is a possible source for the 150-word overlap between the Fusion Memo and the Ecuadorian judgment identified in Dr. Leonard's Example 2.

27. We ran the same analysis for other overlaps in the Fusion Memo, as well as for overlaps between the Ecuadorian judgment and plaintiffs' work product documents known as the Clapp Report, Index Summaries, Fajardo Trust email, Draft Alegato and the Selva Viva Data Compilation. Based on our review, neither the Fusion Memo, the Clapp Report, the Index Summaries, the Fajardo Trust email, the Draft Alegato, nor the Selva Viva Data Compilation appear in the trial court record.

28. We then reexamined the individual cited instances of overlap or similarity that Dr. Leonard determined would indicate plagiarism. Aside from overlaps attributable to direct quotations or titles of specific documents, our analysis confirmed that none of the nine individual instances of substantial textual overlap with the Ecuadorian judgment identified by Dr. Leonard were found in the trial court record. In my opinion, many of these overlaps (e.g. Leonard Example 2) would be sufficient *by themselves* to indicate plagiarism.

29. The files we received were electronic copies which had been subjected to optical character recognition ("OCR"), which is a process by which hard copies are scanned and processed to create electronic files that can be viewed on the computer. In the abstract, poor quality OCR can reduce performance of computer-based text analysis generally, although the amount of performance reduction varies with the type of analysis performed, with the quality of the image, and with the quality of the OCR engine used.

30. Although our original analysis attempted to control for this, we ran additional analyses to determine the effects of OCR quality. To do this, we compared all of the documents in the court record with a "standard corpus" of well-curated Spanish. To obtain this corpus, we harvested Spanish Wikipedia articles, the MultiUN Spanish Corpus, and in addition the Spanish version of the Google N-Gram Corpus (from 1999-2010) to make a baseline histogram of all Spanish words and their frequencies as well as the frequencies of the characters and character n-grams that comprise them.

31. Each document in the trial court record was then compared to determine how close its distribution was to the baseline, and by extension how likely it was to be reasonable Spanish. A document that conforms in all respects to the Spanish corpus is presumed to be perfect Spanish in perfect condition. The extent of the difference is a measure both of the quality of writing and the quality of OCR. (E. g., a bad OCR will produce many errors, but a badly-written document or an unusual type of document such as a table, a map, or a mineralogy report would also be unusual and divergent, hence this is a conservative analysis.).

32. Based on my review, the overall scanning quality was quite high. My comparative analysis of the documents against the well-curated Spanish corpus indicates that no more than 1-1.5% of the documents in the court record were unsearchable. The vast majority of the court record was extremely similar to our model of Spanish obtained from machine-readable sources and a relatively few number of outliers had an extraordinary degree of difference.

33. We collected all of these outlier documents and reviewed them by hand. All of these documents were manually examined by at least two staff members of J&A, one of whom was myself. Many of the outlier documents proved, upon manual inspection, to be non-text documents such as photographic images, tables, or maps. This manual review of the outlier

documents failed to identify any that were the Fusion Memo, the Clapp Report, the Index Summaries, the Fajardo Trust email, the Draft Alegato, or the Selva Viva Database, and failed to provide a source for the overlap between the plaintiffs' work product documents and the Ecuadorian judgment.

**Conclusions**

34. With the aid of computers, we have searched the entire trial court record for the plaintiffs' work product documents and for potential sources for the textual overlaps in the Ecuadorian judgment identified by Dr. Leonard. This computer search has covered all possible sources of these texts within the court record. None of the textual overlaps identified by Dr. Leonard are potentially explainable by a source in the court record.

35. For a number of the longer overlaps, we have conducted a more general search looking for any "near-misses" that might have been garbled in process of analysis (such as by the OCR process). We have found no such near misses.

36. An analysis of the statistical properties of the documents in the court record indicates that a relative few of the documents could not be computer analyzed. I (as well as members of my staff) have personally hand-analyzed these documents and found no source for the overlapping text.

37. Having personally examined all possible sources for these texts by computer, and having personally examined by hand the primary candidates for OCR errors, I have concluded to a reasonable degree of scientific certainty that the Fusion Memo, the Clapp Report, the Index Summaries, the Fajardo Trust email, the Draft Alegato, and the Selva Viva Data Compilation are not in the trial court record. Moreover, I conclude to a reasonable degree of scientific certainty

that the textual overlaps between the Ecuadorian judgment and the plaintiffs' work product documents identified by Dr. Leonard cannot be derived from sources with the trial court record.

I declare under penalty of perjury under the laws of the United States of America that the foregoing is true and correct. Executed on October 9, 2013.

Patrick Juola, Ph. D

JUOLA WITNESS DECLARATION.DOCX