PharmaSUG 2015 - Paper BB13

A Unique Way to Annotate Case Report Forms (CRFs) in PDF, Using Forms Data Format (FDF) Techniques

Boxun Zhang, Seattle Genetics Inc., Bothell, WA
Tyler Kelly, Seattle Genetics Inc., Bothell, WA

ABSTRACT

One of the essential tasks for programmers, as part of the internal processes and/or a component of a regulatory submission, is to annotate CRFs. While there are various approaches to accomplish this, Adobe Acrobat is commonly used. To overcome labor intensiveness of incorporating PDF annotations manually, creating an FDF file provides the possibility of a repository to store and manage the annotations. As these annotations are mapped by page numbers, it's still challenging to automatically assign the annotations back to the CRFs as desired, and across similar studies.

By determining the degree of similarity based on the text strings of CRFs, it would be possible to establish accurate mappings of annotations to CRFs. This paper describes a simple method of using SAS® COMPGED function based on fuzzy matching and explores the dynamic possibilities of incorporating PDF annotations in CRFs.

INTRODUCTION

The purpose of annotated CRFs is to aid programmers and reviewers in navigating a clinical database. It documents the location of the variables included in the submitted datasets on the CRFs. Manually creating and maintaining annotated CRFs is a labor intensive process. Not only does it need human input, but it also usually requires the re-use of the annotations when the CRFs are revised to a new version. Using an FDF file streamlines this process and it saves a tremendous amount of time and effort. As discussed in previous papers [1, 2], adding annotations has been accomplished in a few different ways. Our goal is to expound upon those ideas using third party software and in-house software development.

In an FDF file, the annotations are assigned to the pages where they are located. When they are imported to a new version of CRFs, if any change, the page numbers should be properly re-assigned. In another scenario, which is common either in a research institute or in a pharmaceutical company, most of the annotations can be re-used for similar studies. This is not surprising as the blank CRFs are usually designed in a standard format. The annotations on a single standard CRF can be applied to the same one for a similar study. But in either case, a visual comparison between two version CRFs is still needed and we want to automate this process.

We have found that using Ghostscript in addition to Adobe Acrobat has produced favorable results when tethered to our application. Our paper will describe the methods for using Adobe Acrobat, Ghostscript, the SAS® COMPGED function and our in-house application. We will also describe the workflow for our process and how its benefit can be utilized in other ways.

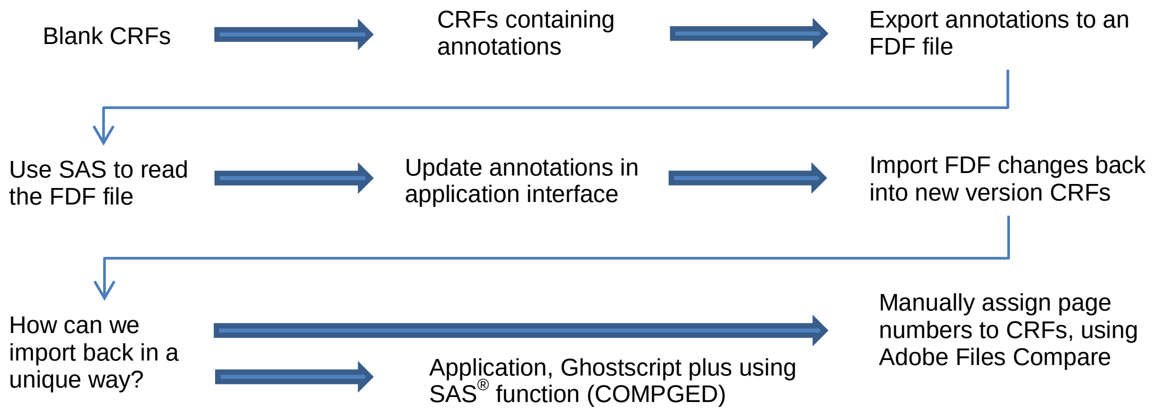
OUR PROJECT

When there are one or more similar studies, especially for clinical trials, they may have some common conventions or standards to follow. The most obvious is the CRFs design. In the scope of the study design, the data management group designs the standard CRFs and uses them for similar studies. That provides SAS® programmers an opportunity to re-use the annotations for the standard CRFs across these studies. We noticed that the SAS® programmers in our group tended to create annotations from scratch when a new study started. We wanted to find a way to create the annotations efficiently using the existing annotations from other similar studies as much as possible.

We have a working solution for re-using annotations in our current annotated CRFs, but the process for making changes can become time consuming and cumbersome. Working together, the idea of our project, CRAFT (CRF Annotation Front-End Tool), was formed. Having a better solution on hand, developed in-house, was the easiest way we could enhance the process of adding/manipulating annotations. When you have a problem, you try to fix it. When you have a working solution, you try to make it faster, better. That is the core idea behind CRAFT.

WORKFLOW

Our workflow is taken from the typical way for updating and analyzing CRF annotations. With our approach, we've developed the application to make FDF annotation updates, compare the results and utilize the SAS® COMPGED function to streamline the process.


Blank CRFs
CRFs containing annotations
Export annotations to an FDF file
Use SAS to read the FDF file
Update annotations in application interface
Import FDF changes back into new version CRFs
How can we import back in a unique way?
Manually assign page numbers to CRFs, using Adobe Files Compare
Application, Ghostscript plus using SAS® function (COMPGED)

OUR SOLUTION

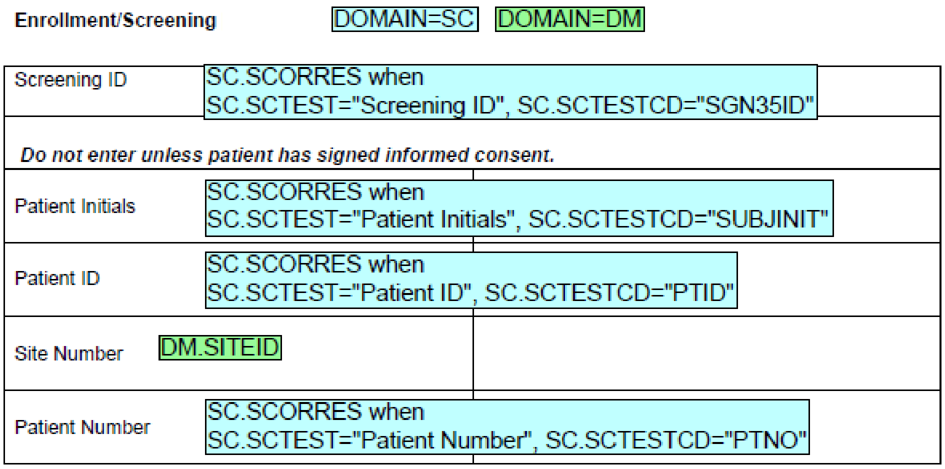
The first challenge for us is to read the text strings from CRFs (PDF files) in SAS®. The purpose is to establish accurate mappings of annotations to CRFs, based on the relationship of the relevant text strings. As described in Wooding's paper [3], Ghostscript can be used to convert PDF files to flat files, e.g., ASCII files. Using SAS®, you can read in ASCII files and build SAS® datasets. Converting PDF files to ASCII files can be done by calling a batch file through SAS® "X" command. In order to use the batch file, we need to generate the post script file for converting PDF files before calling Ghostscript. This can be easily done in Adobe Acrobat. In the Windows environment, the command could change to a different directory, run an executable file or run a batch file. The null step is:

```
DATA _NULL_;
  X “CD H:\ACRFS”;
  X “CREATEFLAT.BAT”; /* PS2ASCII INPUTPS OUTTXT */
RUN;
```

Within the batch file, we can choose several settings for the Ghostscript package `ps2ascii.ps`, (see below). As the batch file can be written and generated within SAS®, we can pass these parameters to tell SAS® which file to read and where to store the output.

```
cd “C:\Program Files\gs\gs9.15\bin”
gswin64c.exe -sstdout=H:\ACRFS\ascii.txt -dSIMPLE -dNODISPLAY -dDELAYBIND
-dWRITESYSTEMDICT -f “C:\Program Files\gs\gs9.15\lib\ps2ascii.ps” H:\ACRFS\crfs.ps
```

Below is an example of the annotated CRFs. The highlighted boxes are the annotations created in Adobe Acrobat. They are properly located on this CRF page. The text file, which is converted by calling Ghostscript, captures all the text strings on this page. What we found is that these text strings are laid out in this order: the strings on CRFs itself come first and then the strings for the annotations second.



Enrollment/Screening DOMAIN=SC DOMAIN=DM

Screening ID	SC.SCORRES when SC.SCTEST="Screening ID", SC.SCTESTCD="SGN35ID"
Do not enter unless patient has signed informed consent.	
Patient Initials	SC.SCORRES when SC.SCTEST="Patient Initials", SC.SCTESTCD="SUBJINIT"
Patient ID	SC.SCORRES when SC.SCTEST="Patient ID", SC.SCTESTCD="PTID"
Site Number DM.SITEID	
Patient Number	SC.SCORRES when SC.SCTEST="Patient Number", SC.SCTESTCD="PTNO"

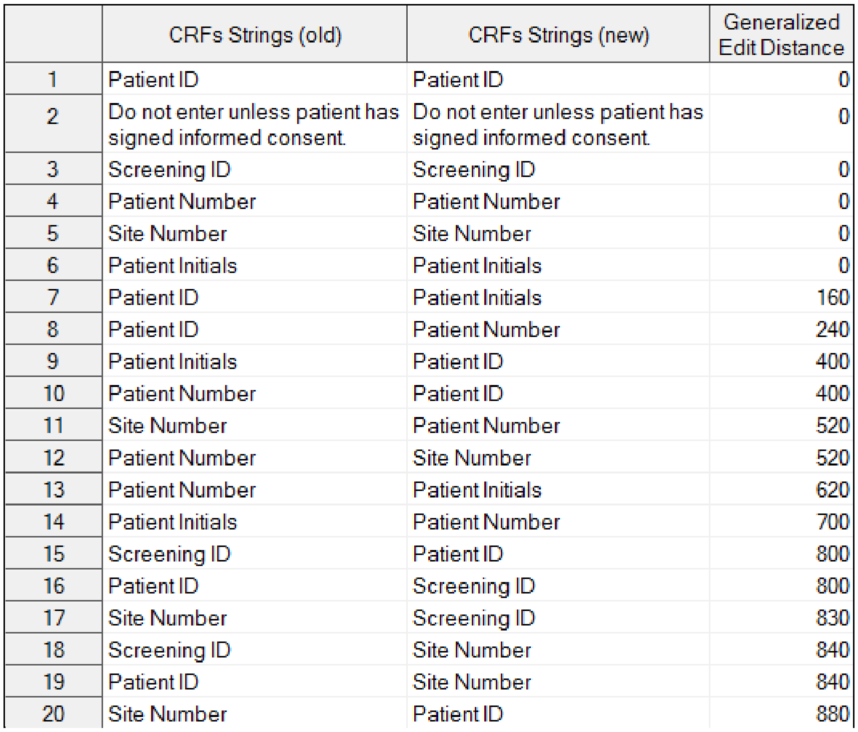
```
Enrollment/Screening
Screening ID
Do not enter unless patient has signed informed consent.
Patient Initials
Patient ID
Site Number
Patient Number

DOMAIN=SC DOMAIN=DM
SC.SCORRES when SC.SCTEST="Screening ID", SC.SCTESTCD="SGN35ID"
SC.SCORRES when SC.SCTEST="Patient ID", SC.SCTESTCD="PTID"
DM.SITEID
SC.SCORRES when SC.SCTEST="Patient Number", SC.SCTESTCD="PTNO"
SC.SCORRES when SC.SCTEST="Patient Initials", SC.SCTESTCD="SUBJINIT"
```

Output 1. Text file generated by Ghostscript

In Output 1, the top section contains the strings from CRFs and the bottom section captures the annotations in the order of their occurrences in the FDF file. Once we read them in SAS® and put them in a single dataset, we can have a good binding for the annotations to a particular CRF page. In addition, these annotations belong to that particular page by recognizing the CRFs strings. We call the batch file to convert the new version CRFs and create a separate SAS® dataset for comparison. By using the SAS® COMPGED function to calculate the degree of similarities for each individual datasets, we can find the standard CRFs across both studies. As described in Staum's paper [4], we used this easy method to build a mapping sheet between two sets of CRFs. Let's look into the details. We first make a Cartesian join for these two datasets. Then we select the top 15 minimum, generalized edit distances between two strings. We want to use the sum of these values to be a similarity indicator.

```
proc sql;
  create table compged as
  select *, compged(old,new) as ged
  from CRF1, CRF2
  order by ged;
quit;
```



	CRFs Strings (old)	CRFs Strings (new)	Generalized Edit Distance
1	Patient ID	Patient ID	0
2	Do not enter unless patient has signed informed consent.	Do not enter unless patient has signed informed consent.	0
3	Screening ID	Screening ID	0
4	Patient Number	Patient Number	0
5	Site Number	Site Number	0
6	Patient Initials	Patient Initials	0
7	Patient ID	Patient Initials	160
8	Patient ID	Patient Number	240
9	Patient Initials	Patient ID	400
10	Patient Number	Patient ID	400
11	Site Number	Patient Number	520
12	Patient Number	Site Number	520
13	Patient Number	Patient Initials	620
14	Patient Initials	Patient Number	700
15	Screening ID	Patient ID	800
16	Patient ID	Screening ID	800
17	Site Number	Screening ID	830
18	Screening ID	Site Number	840
19	Patient ID	Site Number	840
20	Site Number	Patient ID	880

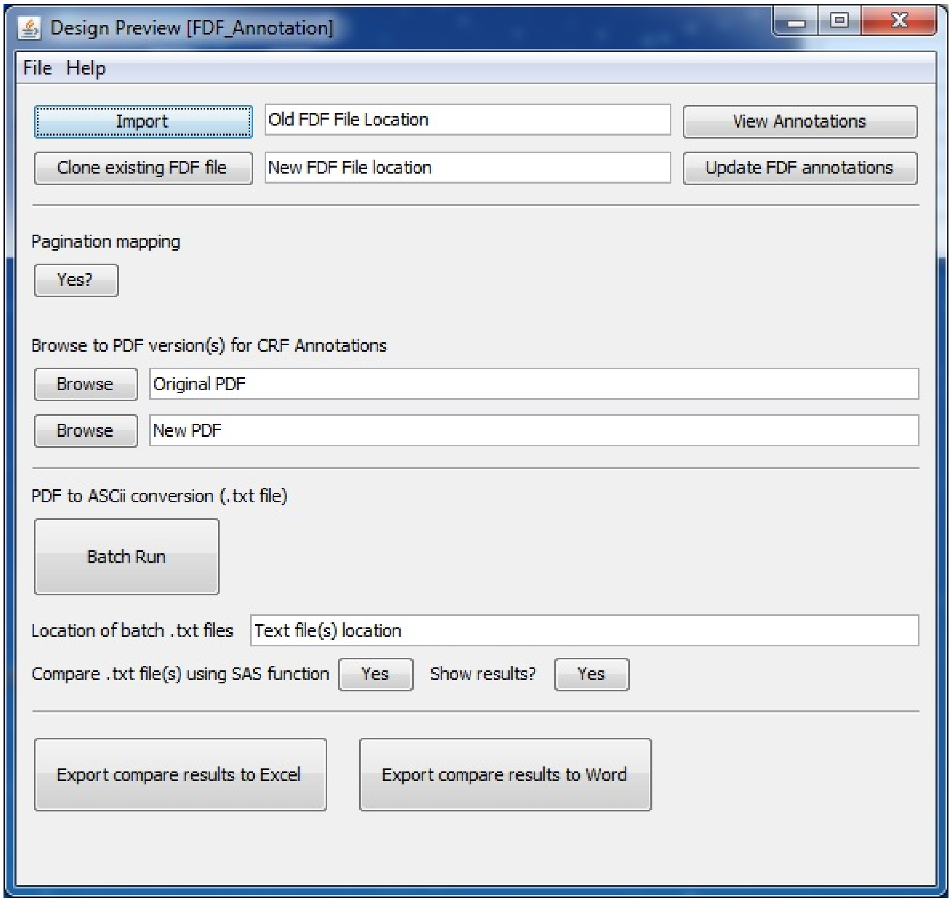
Based on the calculation of the top minimum generalized edit distance values, we can determine the most similar datasets by looking for the smallest value. Without a doubt, two distinct forms will produce a huge value for the generalized edit distance. At this point, we can figure out the relationship between two version CRFs or two sets of CRFs for the similar studies and we can re-assign the new page numbers. This process has been automated within our in-house application.

OUR IN-HOUSE APPLICATION

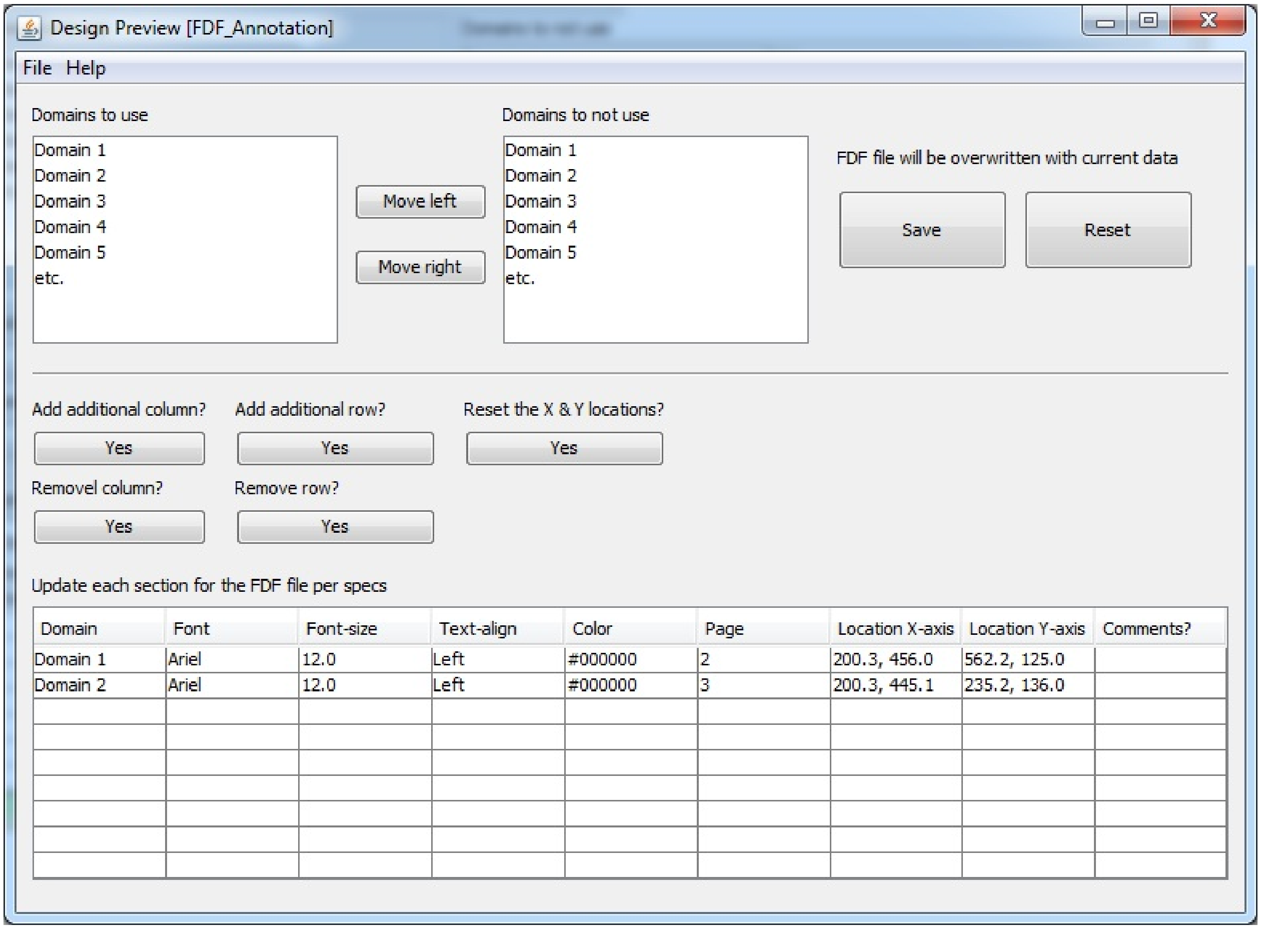
Our in-house application software is designed to streamline the workflow, see Display 1. Putting all things together, we can complete the steps by clicking the buttons. It mainly supports five functions:

- Call SAS$^{®}$ to read in an FDF file and build the SAS$^{®}$ dataset on the host server to store the annotation's attributes.
- Display the annotations in a user friendly interface and provide a direct edit tool to modify the attributes or to add a new annotation on a particular page.
- Call SAS$^{®}$ to generate a new FDF file based on the modified/new annotations from the SAS$^{®}$ dataset.
- Call Ghostscript batch file to convert PDF files, create individual SAS$^{®}$ datasets, run SAS$^{®}$ COMPGED function to calculate generalized edit distance and determine the similar CRFs for matching.
- Export comparison results to an Excel/Word file and automatically assign the new page numbers to the new annotations based on the accurate mappings described earlier.

In the annotation edit interface, see Display 2, we can edit many attributes, such as font, font size, color, the x-axis/y-axis location and page number. In addition, we can edit and cross check the annotations by domain, so that we would efficiently navigate and handle the relevant annotations in one place. It is also doable to add or remove an annotation in this interface.


Design Preview [FDF_Annotation]
File
Help
Import
Old FDF File Location
View Annotations
Clone existing FDF file
New FDF File location
Update FDF annotations
Pagination mapping
Yes?
Browse to PDF version(s) for CRF Annotations
Browse
Original PDF
Browse
New PDF
PDF to ASCii conversion (.txt file)
Batch Run
Location of batch .txt files
Text file(s) location
Compare .txt file(s) using SAS function
Yes
Show results?
Yes
Export compare results to Excel
Export compare results to Word

Display 1. CRAFT: Workflow Console



Domains to use

Domain 1
Domain 2
Domain 3
Domain 4
Domain 5
etc.

Domains to not use

Domain 1
Domain 2
Domain 3
Domain 4
Domain 5
etc.

FDF file will be overwritten with current data

Add additional column?

Add additional row?

Reset the X & Y locations?

Removel column?

Remove row?

Update each section for the FDF file per specs

Domain	Font	Font-size	Text-align	Color	Page	Location X-axis	Location Y-axis	Comments?
Domain 1	Ariel	12.0	Left	#000000	2	200.3, 456.0	562.2, 125.0	
Domain 2	Ariel	12.0	Left	#000000	3	200.3, 445.1	235.2, 136.0	

Display 2. CRAFT: Annotation Edit Interface

CONCLUSION

As described in this paper, we used different tools to find the solution and accomplished our project. This method works well for the annotation creation, especially for the similar CRFs. Based on the fuzzy matching, we explored the usage for the SAS® COMPGED function and that's a good example of how we solved our problem by using SAS® base functions.

REFERENCES

- [1] Spruck, Dirk and Kawohl, Monika. 2004. "Using SAS to Speed Up Annotating Case Report Forms in PDF Format." *Proceedings of the 2004 Pharmaceutical Industry SAS® Users Group Conference*. Available at http://www.lexjansen.com/pharmasug/2004/coderscorner/cc02.pdf
- [2] Hufford, Walter. 2014. "Automating Production of the blankcrf.pdf." *Proceedings of the 2014 Pharmaceutical Industry SAS® Users Group Conference*. Available at http://www.pharmasug.org/proceedings/2014/CC/PharmaSUG-2014-CC21.pdf
- [3] Wooding, Nat. 2005. "EXTRACTING DATA FROM PDF FILES." *Proceedings of the 2005 Southeast SAS® Users Group Conference*. Available at http://analytics.ncsu.edu/sesug/2005/SER10_05.PDF
- [4] Staum, Paulette. 2007. "Fuzzy Matching using the COMPGED Function." *Proceedings of the 2007 Northeast SAS® Users Group Conference*. Available at http://www.nesug.org/Proceedings/nesug07/ap/ap23.pdf

ACKNOWLEDGMENTS

We would like to thank Rajeev Karanam for encouraging us to write this paper and our manager Shefalica Chand for reviewing our paper and assisting us with this project.

RECOMMENDED READING

- PDF Reference, available at http://www.adobe.com/content/dam/Adobe/en/devnet/acrobat/pdfs/pdf_reference_1-7.pdf
- Ghostscript, available at http://www.ghostscript.com/doc/9.15/Readme.htm
- The Little SAS Book 4th/5th Edition(s)

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Boxun Zhang
Seattle Genetics, Inc.
21823 30th Drive Southeast
Bothell, WA 98021
E-mail: bzhang@seagen.com

Tyler Kelly
Seattle Genetics, Inc.
21823 30th Drive Southeast
Bothell, WA 98021
E-mail: tkelly@seagen.com