

UNITED STATES PATENT AND TRADEMARK OFFICE
BEFORE THE PATENT TRIAL AND APPEAL BOARD

GUARDANT HEALTH, INC.,)
)
 Petitioner,)
)
 vs.)
)
 UNIVERSITY OF WASHINGTON)
 THROUGH ITS CENTER FOR)
 COMMERCIALIZATION,)
)
 Patent Owner.)
)

Case No. IPR2022-00817
Patent No. 10,760,127

DEPOSITION OF ANDRES LARREA, Ph.D.

9:06 A.M. PST
TUESDAY, MAY 9, 2023

701 FIFTH AVENUE, SUITE 5100
SEATTLE, WASHINGTON

REPORTED BY: VICKY L. PINSON, RPR-CCR WASHINGTON 2559
CALIFORNIA NO. 9845; OREGON NO. 16-0442

Page 1

A P P E A R A N C E S

FOR THE PETITIONER AND WITNESS:

JAD A. MILLS
SONJA GERRARD
Wilson Sonsini Goodrich & Rosati
701 Fifth Avenue, Ste 5100
Seattle, WA 98104-7036
206.883.2500
jmills@wsgr.com
sgerrard@wsgr.com

FOR THE PATENT OWNER:

R. WILSON "TREY" POWERS III, Ph.D.
DAVID H. HOLMAN, PH.D.
Sterne, Kessler, Goldstein & Fox, PLLC
1100 New York Avenue NW
Washington, DC 20005
202.371.2600
tpowers@sternekessler.com
dholman@sternekessler.com

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25

I N D E X

PAGE

Cross-Examination by Mr. Powers	4
Redirect Examination by Mr. Mills	59
Recross-Examination by Mr. Powers	61

EXHIBITS FOR IDENTIFICATION PAGE

Exhibit 1	Table of Contents	14
Exhibit 2	SMRTbell template	28
Exhibit 3	Schematic of linear single-stranded RCR product generated from a polymerase going around a SMRTbell three times	30
Exhibit 4	Schematic of linear single-stranded RCR product generated from a polymerase going around a SMRTbell four times	31
Exhibit 5	Schematic of linear single-stranded RCR product generated from a polymerase going around a SMRTbell product ten times	31
Exhibit 6	Pacific Biosciences Glossary of Terms	39
Exhibit 7	CHROMA SPIN Columns User Manual	49
Exhibit 8	Sensors and Actuators B: Chemical	51
Exhibit 9	U.S. Patent Application No. US 2009/0024331A1	57

1 Seattle, Washington; May 9, 2023

2 9:06 a.m. PST

3 * * *

4
5 ANDRES LARREA, Ph.D.,

6 remotely sworn as a witness by the Certified Court

7 Reporter, testified as follows:

8
9 CROSS-EXAMINATION

10 BY MR. POWERS:

11 Q. Good morning, Dr. Larrea. 09:06AM

12 A. Good morning. 09:06AM

13 Q. I'm Ralph Powers, III from Sterne, Kessler, 09:06AM
14 Goldstein & Fox on behalf of The University of 09:06AM
15 Washington. With me in the room is David Holman, also 09:06AM
16 from Stern Kessler. 09:06AM

17 Have you been deposed before? 09:06AM

18 A. Yes. I've been deposed once before. 09:06AM

19 Q. When was that? 09:06AM

20 A. I don't recall the exact time. 09:06AM

21 Q. Do you recall the approximate time? 09:07AM

22 A. I'd say approximately 2015. 09:07AM

23 Q. And did you serve as an expert witness in that 09:07AM
24 case? 09:07AM

25 A. No, I did not. 09:07AM

Page 4

1 Q. You served as a fact witness in that case. Is 09:07AM
2 that correct? 09:07AM
3 A. Yes, I believe that is correct. 09:07AM
4 Q. What was the nature of that case? 09:07AM
5 A. It was a lawsuit. 09:07AM
6 Q. Okay. Can you be a little bit more specific? 09:07AM
7 A. It regarded an accident at my house. 09:07AM
8 Q. Was it a negligence lawsuit? 09:07AM
9 A. It was a negligence lawsuit. 09:07AM
10 Q. Were you the defendant? 09:07AM
11 A. I was not the defendant. 09:07AM
12 Q. Were you the plaintiff? 09:07AM
13 A. Yes. 09:07AM
14 Q. Who was the defendant in that lawsuit? 09:07AM
15 A. Home Depot, I believe. 09:08AM
16 Q. What was the nature of the accident? 09:08AM
17 A. I was pushing on a box that they were 09:08AM
18 delivering, and I tore my Achilles. 09:08AM
19 Q. I hope you're okay now? 09:08AM
20 A. I am. I've since torn the other Achilles in 09:08AM
21 an unrelated event. 09:08AM
22 Q. Okay. So since you've been deposed before, 09:08AM
23 you may know this, but I just want to go over some 09:08AM
24 ground rules with you. 09:08AM
25 First, as you can see, this is all being 09:08AM

1 transcribed so we'll need to get verbal answers from 09:08AM
2 you. Do you understand that? 09:08AM
3 A. Yes, I do. 09:08AM
4 Q. Okay. Let me know if you don't understand a 09:08AM
5 question; but if you go ahead and answer it, I'll 09:08AM
6 assume that you understood the question. 09:08AM
7 Is that okay? 09:08AM
8 A. That's okay. 09:08AM
9 Q. And it's also important that we try not to 09:08AM
10 talk over each other so that the court reporter can do 09:09AM
11 her job. 09:09AM
12 Do you understand? 09:09AM
13 A. I do. 09:09AM
14 Q. Okay. You're also under oath, so you need to 09:09AM
15 give full, complete and accurate answers. 09:09AM
16 Is that acceptable? 09:09AM
17 A. Yes, it is. 09:09AM
18 Q. Your counsel may object, but you still need to 09:09AM
19 answer the question unless you're instructed not to. 09:09AM
20 Do you understand? 09:09AM
21 A. I do. 09:09AM
22 Q. Okay. I see that you have some documents in 09:09AM
23 front of you. What are those documents? 09:09AM
24 A. These are the documents from -- that include 09:09AM
25 my declaration as well as the documents referenced. 09:09AM

1 Q. Referenced where? 09:09AM

2 A. In the declaration. 09:09AM

3 Q. So that's your -- so you brought into this 09:09AM

4 deposition room a copy of your declaration and copies 09:09AM

5 of all the exhibits cited in your declaration. Is that 09:09AM

6 correct? 09:09AM

7 A. Perhaps not all the exhibits, but most of the 09:09AM

8 exhibits. 09:10AM

9 Q. Some of the exhibits cited in your 09:10AM

10 declaration? 09:10AM

11 A. That is correct. 09:10AM

12 Q. Potentially all of the exhibits cited in your 09:10AM

13 declaration? 09:10AM

14 A. Potentially all. 09:10AM

15 Q. Who selected the documents that you have in 09:10AM

16 front of you? 09:10AM

17 A. I did. 09:10AM

18 Q. And may I inspect those documents? 09:10AM

19 A. (Indicating). 09:10AM

20 Q. Did you make any notes or markings on these 09:10AM

21 documents? 09:10AM

22 A. I did not. 09:10AM

23 Q. I see there's a table of contents at the 09:10AM

24 beginning of this binder. Would you please read out 09:11AM

25 the exhibit numbers that are listed there in the table 09:11AM

1 of contents? 09:11AM

2 A. Yes. 09:11AM

3 Tab 1: Exhibit 1033, Larrea Rebuttal 09:11AM

4 Declaration. 09:11AM

5 Tab 2: Exhibit 1018, US 20090298075A1. 09:11AM

6 Tab 3: Exhibit 1020, US 7,476,503. 09:11AM

7 Tab 4: Exhibit 1021, Travers. 09:11AM

8 Tab 5: Exhibit 1035, Template Preparation 09:11AM

9 Sequencing Guide. 09:11AM

10 Tab 6: Exhibit 1036, Chin 2011. 09:11AM

11 Tab 7: Exhibit 2001, Eid 2009. 09:12AM

12 Tab 8: Exhibit 2002, Niedringhaus 2010. 09:12AM

13 Tab 9: Exhibit 2006, Egan 2012. 09:12AM

14 Tab 10: Exhibit 2014, Satija Declaration. 09:12AM

15 Tab 11: Exhibit 2021, PacBio Technology 09:12AM

16 Backgrounder 2010. 09:12AM

17 Tab 12: Exhibit 2026, Eid Supplemental 09:12AM

18 Material. 09:12AM

19 Q. And you reviewed all of those documents in 09:12AM

20 preparation for your declaration that you filed in this 09:12AM

21 case. Is that correct? 09:12AM

22 A. Yes, that is correct. 09:12AM

23 Q. Did you review the entire Satija declaration? 09:12AM

24 A. I have read the entire Satija declaration. I 09:13AM

25 focused my studies on the problems raised by Satija. 09:13AM

Page 8

1 Q. Did you review all the materials cited in the 09:13AM
2 Satija declaration? 09:13AM

3 A. I did not review all the material. 09:13AM

4 Q. Aside from the documents in front of you that 09:13AM
5 you just listed, did you consider any other materials 09:13AM
6 in rendering your opinions in your declaration in this 09:13AM
7 case? 09:13AM

8 MR. MILLS: Objection to form. 09:13AM

9 A. No. The material that I used in prepping this 09:13AM
10 declaration is here. 09:14AM

11 Q. Did you review the petition filed in this 09:14AM
12 case? 09:14AM

13 A. I did not review the petition filed in this 09:14AM
14 case. My focus was really on the -- Dr. Satija's 09:14AM
15 testimony about the four alleged problems. 09:14AM

16 Q. Did you review Dr. Quackenbush's first 09:14AM
17 declaration filed in this case? 09:14AM

18 A. There were some excerpts in Satija's 09:14AM
19 declaration, but I did not do that. I really focused 09:14AM
20 on Dr. Satija's testimony regarding the four alleged 09:14AM
21 problems. 09:14AM

22 Q. Did you review Dr. Quackenbush's second 09:14AM
23 declaration filed in this case? 09:14AM

24 A. I did not. Again, I really focused on the 09:14AM
25 testimony regarding the four alleged problems. 09:15AM

1 Q. If you go back to your table of contents 09:15AM
2 there -- sorry -- the table of contents for the binder, 09:15AM
3 who selected those documents for you to review? 09:15AM
4 MR. MILLS: Objection to form. 09:15AM
5 A. I'm not sure that I understand your question. 09:15AM
6 Q. Let me ask it a little differently. 09:15AM
7 Who selected those documents for you to review 09:15AM
8 in preparation of your declaration filed in this case? 09:15AM
9 A. If you mean specifically for the declaration 09:15AM
10 today, then, like I said, I selected the documents that 09:16AM
11 were to be brought in today. 09:16AM
12 Q. Today is what we call a deposition. But you 09:16AM
13 also filed a declaration; right? 09:16AM
14 A. Um-hmm. 09:16AM
15 Q. Who selected -- and you said that those 09:16AM
16 documents in your table of contents, which we'll mark 09:16AM
17 as Larrea Exhibit 1, you also relied on those documents 09:16AM
18 in preparation of your declaration. Right? 09:16AM
19 A. Yes. That is correct. 09:16AM
20 Q. And who selected those documents for you to 09:16AM
21 rely on in preparing your declaration in this case? 09:16AM
22 MR. MILLS: Objection to form. 09:16AM
23 A. Many of these exhibits were taken from 09:16AM
24 Satija's declaration, which I was asked to review. 09:17AM
25 Q. Okay. And did counsel provide you with some 09:17AM

1 of those documents? 09:17AM

2 A. I spent time working at PacBio, and I was 09:17AM

3 familiar with many of these, so I wasn't provided 09:18AM

4 anything that I had not previously read. 09:18AM

5 Q. That doesn't quite answer my question. 09:18AM

6 Did counsel provide you with the documents 09:18AM

7 that are listed in your table of contents? 09:18AM

8 MR. MILLS: Objection to form. Asked and 09:18AM

9 answered. 09:18AM

10 A. I guess I don't understand exactly what you're 09:18AM

11 asking. 09:18AM

12 Q. In your preparation and your work for this 09:18AM

13 case, did counsel provide you the documents listed in 09:18AM

14 your table of contents there? 09:18AM

15 MR. MILLS: Same objections. 09:18AM

16 A. I reviewed the Satija declaration. I 09:19AM

17 identified certain problems, and we reviewed and 09:19AM

18 identified the relevant material to address those. 09:19AM

19 Q. And as part of that, did counsel provide you 09:19AM

20 with the documents listed in your table of contents 09:19AM

21 there? 09:19AM

22 MR. MILLS: Same objections. 09:19AM

23 A. They did not provide me with anything that I 09:19AM

24 had not previously read. 09:19AM

25 Q. Okay, that really doesn't answer my question, 09:19AM

1 though. Did counsel provide you with the documents 09:20AM
2 that are listed in your table of contents? 09:20AM
3 MR. MILLS: Same objections. 09:20AM
4 A. (Pause). Counsel highlighted some -- 09:20AM
5 MR. MILLS: So -- 09:20AM
6 MR. POWERS: Please let the witness 09:21AM
7 finish. 09:21AM
8 MR. MILLS: No, I'm going to register an 09:21AM
9 objection. 09:21AM
10 MR. POWERS: That's fine. 09:21AM
11 MR. MILLS: Yeah, thank you. 09:21AM
12 MR. POWERS: Testimony proceeds. 09:21AM
13 MR. MILLS: Yeah, my objection is on the 09:21AM
14 basis of work product and privilege, so I'm going to 09:21AM
15 issue a cautionary instruction. 09:21AM
16 To the extent that the answer -- the question 09:21AM
17 calls for the disclosure of any confidential 09:21AM
18 conversations with counsel that do not form a basis for 09:21AM
19 your opinions, I would instruct you not to answer the 09:21AM
20 question. 09:21AM
21 Q. (By Mr. Powers) Yeah, I'm not calling for any 09:21AM
22 privileged information. 09:21AM
23 The question is simply: Did counsel provide 09:21AM
24 you with the documents there listed in your table of 09:21AM
25 contents? 09:21AM

1 A. No documents in my declaration and no part of 09:21AM
2 my analysis was based on material that I had not 09:22AM
3 previously seen. 09:22AM

4 Q. Okay, fair enough. 09:22AM

5 Did counsel provide you with the documents 09:22AM
6 that are listed in your table of contents? 09:22AM

7 MR. MILLS: Objection to form. Asked and 09:22AM
8 answered. 09:22AM

9 A. For example, the template preparation 09:22AM
10 sequencing guide I would hand out to people as I was 09:22AM
11 training them many years ago. Similar with Travers; 09:22AM
12 similar with Chin; similar with Eid. So there was no 09:23AM
13 material in here that was new to me. 09:23AM

14 Q. Okay. But as part of your work in preparing 09:23AM
15 your declaration in this case, counsel provided you 09:23AM
16 with those documents. Is that fair? 09:23AM

17 A. I would not say that that is an accurate 09:23AM
18 description. 09:23AM

19 Q. What would you say is an accurate description? 09:23AM

20 A. I reviewed Dr. Satija's declaration, I offered 09:24AM
21 comments on that, and then those comments were 09:24AM
22 memorialized into the declaration. 09:24AM

23 Q. And who wrote your declaration? 09:24AM

24 A. Like I said, it was -- I reviewed his, 09:24AM
25 Dr. Satija's declaration. I made comments on that 09:24AM

1 declaration. And then the exact writing was a 09:24AM
2 collaborative effort. 09:25AM
3 Q. Collaborative effort among whom? 09:25AM
4 A. Among myself and counsel. 09:25AM
5 Q. Let's go ahead and mark that table of contents 09:25AM
6 as Larrea Exhibit 1. 09:25AM
7 (Exhibit 1 was marked for identification.) 09:25AM
8 Q. (By Mr. Powers) Did you review the patent at 09:25AM
9 issue in this case? 09:25AM
10 A. I did not review the patent. I really focused 09:25AM
11 my work on the problems in Dr. Satija's declaration. 09:26AM
12 Q. Did you review any prosecution history 09:26AM
13 documents as part of your work for this case? 09:26AM
14 A. I did not review any prosecution document. I 09:26AM
15 really focused on the four problems in Dr. Satija's 09:26AM
16 declaration. 09:26AM
17 Q. Aside from counsel, did you speak with anyone 09:26AM
18 in arriving at your opinion expressed in your 09:26AM
19 declaration? 09:26AM
20 A. No, I did not. 09:26AM
21 Q. And you already told me that you did not 09:27AM
22 consider any references that you did not cite in your 09:27AM
23 declaration. Is that correct? 09:27AM
24 MR. MILLS: Objection to form. 09:27AM
25 Q. (By Mr. Powers) There's a double negative 09:27AM

1	declaration regarding the four problems that are not	09:29AM
2	addressed in your declaration. Is that correct?	09:29AM
3	MR. MILLS: Objection to form.	09:29AM
4	A. There may be. I don't recall.	09:29AM
5	Q. (By Mr. Powers) Did you meet with anyone in	09:30AM
6	preparation for your deposition today?	09:30AM
7	A. I did.	09:30AM
8	Q. Who did you meet with?	09:30AM
9	A. I met with counsel.	09:30AM
10	Q. Can you name them?	09:30AM
11	A. I met with Jad, Sonja, and Mike.	09:30AM
12	Q. And about how long did you meet with counsel?	09:30AM
13	A. We met the first time for about an hour.	09:30AM
14	Q. When was the first time?	09:30AM
15	A. I don't recall.	09:30AM
16	Q. Okay. Did you meet again a second time?	09:30AM
17	A. We did.	09:30AM
18	Q. And how long did you meet with them for the	09:30AM
19	second time?	09:30AM
20	A. One hour.	09:30AM
21	Q. Did you meet a third time?	09:30AM
22	A. We did.	09:30AM
23	Q. How long did you meet for the third time?	09:30AM
24	A. About a day.	09:31AM
25	Q. And was that day yesterday?	09:31AM

1	A. That day was yesterday.	09:31AM
2	Q. So about eight hours yesterday?	09:31AM
3	A. That sounds about right.	09:31AM
4	Q. So all in all, you met with counsel for about	09:31AM
5	ten hours. Is that correct?	09:31AM
6	A. That is correct.	09:31AM
7	Q. Can you go to paragraph 6 of your declaration?	09:31AM
8	It says that you: Introduced next-generation	09:31AM
9	sequencing technologies to improve efficiencies and cut	09:31AM
10	costs.	09:31AM
11	Do you see that?	09:31AM
12	A. I do.	09:31AM
13	Q. Is efficiency important to a skilled artisan?	09:31AM
14	MR. MILLS: Objection to form. Scope.	09:31AM
15	A. In that particular experiment it was.	09:32AM
16	Q. (By Mr. Powers) Is efficiency generally	09:32AM
17	important to a skilled artisan?	09:32AM
18	MR. MILLS: Same objection.	09:32AM
19	Q. (By Mr. Powers) In the context of next-gen	09:32AM
20	sequencing?	09:32AM
21	A. I would say it depends.	09:32AM
22	Q. What's it depend on?	09:32AM
23	A. It would depend on the goal of the experiment.	09:32AM
24	Q. All other things being equal, isn't it true	09:32AM
25	that a skilled artisan would appreciate a more	09:32AM

Page 17

1 efficient approach -- 09:32AM

2 MR. MILLS: Objection to form. 09:32AM

3 Q. (By Mr. Powers) -- in next generation 09:33AM

4 sequencing technology? 09:33AM

5 MR. MILLS: Objection to form and scope. 09:33AM

6 A. All other things are rarely equal, and so I 09:33AM

7 would say it depends on the experimental conditions. 09:33AM

8 Q. Efficiency is a goal in next-generation 09:33AM

9 sequencing, though; correct? 09:33AM

10 MR. MILLS: Objection to form. Scope. 09:33AM

11 A. Efficiency is one parameter in next-gen 09:33AM

12 sequencing. 09:33AM

13 Q. (By Mr. Powers) And so skilled artisans are 09:33AM

14 looking to be as efficient as possible in accomplishing 09:34AM

15 their goals in next-generation sequencing. Correct? 09:34AM

16 MR. MILLS: Object to form. And outside 09:34AM

17 the scope. 09:34AM

18 A. It depends a great deal upon the exact 09:34AM

19 experimental design in question. 09:34AM

20 Q. (By Mr. Powers) But we agreed that efficiency 09:34AM

21 is important for artisans in next-generation 09:34AM

22 sequencing. Correct? 09:34AM

23 MR. MILLS: Objection to form. Scope. 09:34AM

24 A. There are many experiments in which efficiency 09:34AM

25 is not the primary goal of the process. It will depend 09:34AM

1 on the question and the design. 09:34AM

2 Q. (By Mr. Powers) And you say that you 09:34AM

3 introduce next-generation sequencing technologies to 09:35AM

4 the lab to improve efficiency and cut costs. 09:35AM

5 What specifically did you do to improve 09:35AM

6 efficiency and cut costs? 09:35AM

7 A. In that particular example? 09:35AM

8 Q. Yes. 09:35AM

9 A. So for that particular example it was based on 09:35AM

10 reporter cassettes. And next-gen sequencing allowed us 09:35AM

11 to look at replication across the entire genome. 09:35AM

12 Q. And so how did you improve efficiency? 09:35AM

13 A. Because I was looking at it across the entire 09:35AM

14 genome rather than a series of cassettes that would 09:35AM

15 have to be designed and inserted into different parts 09:35AM

16 of the genome. 09:35AM

17 Q. So you increased the throughput. Is that 09:35AM

18 right? 09:36AM

19 A. There was an increase in throughput also. 09:36AM

20 Q. And was that because of the adoption of 09:36AM

21 next-gen sequencing as opposed to older sequencing 09:36AM

22 technology? 09:36AM

23 A. In that particular experiment it was related 09:36AM

24 to next-gen sequencing. 09:36AM

25 Q. And so that was an important goal in that 09:36AM

1 experiment to improve efficiency. Correct? 09:36AM

2 A. In that experiment, it was. 09:36AM

3 Q. And it was also an important goal to cut 09:36AM

4 costs. Correct? 09:36AM

5 A. The cost-cutting was less important to that 09:36AM

6 particular PI. 09:36AM

7 Q. Were they Hughes-funded? 09:36AM

8 A. It was at the NIH. 09:36AM

9 Q. I see. Do you agree, though, that skilled 09:37AM

10 artisans are generally motivated to improve efficiency 09:37AM

11 and cut costs? 09:37AM

12 MR. MILLS: Objection to form. Scope. 09:37AM

13 A. Again, it depends a great deal on the exact 09:37AM

14 experiment. 09:37AM

15 Q. (By Mr. Powers) If there are two approaches 09:37AM

16 and one is more efficient than the other and they yield 09:37AM

17 comparable results, do you agree that a skilled artisan 09:37AM

18 would choose the more efficient approach? 09:37AM

19 MR. MILLS: Objection to form. Scope. 09:37AM

20 A. There are commonly tradeoffs, and so it will 09:37AM

21 depend on the exact experimental goal. 09:38AM

22 Q. (By Mr. Powers) If the goal is achieved under 09:38AM

23 two different approaches, and one approach is more 09:38AM

24 efficient than the other, would the skilled artisan be 09:38AM

25 motivated to choose the more efficient approach? 09:38AM

1 MR. MILLS: Objection to form. Scope. 09:38AM

2 A. It depends a great deal on the specifics of 09:38AM

3 the experiment. 09:39AM

4 Q. (By Mr. Powers) Well, I'm asking you a 09:39AM

5 hypothetical. The goal is achieved, same goal is 09:39AM

6 achieved under two different approaches, but one 09:39AM

7 approach saves costs and is more efficient. 09:39AM

8 Would the skilled artisan be motivated to 09:39AM

9 choose that approach? 09:39AM

10 MR. MILLS: Objection to form. Scope. 09:39AM

11 Asked and answered. 09:39AM

12 A. In this hypothetical there are a number of 09:39AM

13 other factors that would play into it, and so I can't 09:39AM

14 answer on it based on the details that you're giving 09:39AM

15 me. 09:40AM

16 Q. (By Mr. Powers) What other details would you 09:40AM

17 need? 09:40AM

18 MR. MILLS: Objection. Foundation. 09:40AM

19 A. For example, does the inefficient process 09:40AM

20 provide any other advantage? 09:40AM

21 Q. (By Mr. Powers) Let's say the process does 09:40AM

22 not provide any other advantage. It accomplishes the 09:40AM

23 same goals, but one process is more efficient and saves 09:41AM

24 costs. 09:41AM

25 Would the skilled artisan choose the more 09:41AM

1 efficient and less-costly method? 09:41AM

2 MR. MILLS: Objection to form. Outside 09:41AM

3 the scope. 09:41AM

4 A. It would also depend on how costs are 09:41AM

5 calculated. 09:41AM

6 Q. (By Mr. Powers) By any metric in this 09:41AM

7 hypothetical, one method saves costs and is more 09:41AM

8 efficient and the other does not save costs and is less 09:41AM

9 efficient. They achieve equivalent goals. 09:41AM

10 Would a skilled artisan be motivated to choose 09:41AM

11 an approach that is more efficient and saves costs? 09:41AM

12 MR. MILLS: Objection to form. Outside 09:42AM

13 the scope. 09:42AM

14 A. The challenge is that I can think of many 09:42AM

15 cases in which efficiency and cost are not the only 09:42AM

16 drivers. And so the only thing I can tell you is that 09:42AM

17 it is -- it will depend a great deal on the exact 09:42AM

18 experimental design. 09:42AM

19 Q. So sitting here today, you can't say whether 09:42AM

20 or not a person of ordinary skill would choose of two 09:42AM

21 comparable approaches, one that is cheaper and more 09:42AM

22 efficient versus one that is more expensive and less 09:42AM

23 efficient. 09:42AM

24 You couldn't say? 09:42AM

25 MR. MILLS: Objection to form. Outside 09:42AM

1 the scope. Asked and answered. 09:42AM

2 A. I'm saying that if -- that efficiency and cost 09:42AM

3 are two factors, and there are many other factors that 09:43AM

4 play into experimental design. 09:43AM

5 Q. But I'm asking you: The goal achieved is the 09:43AM

6 same under either scenario and all other things being 09:43AM

7 equal, will the person of ordinary skill be motivated 09:43AM

8 to choose the less costly and more efficient approach? 09:43AM

9 MR. MILLS: Objection to form. Outside 09:43AM

10 the scope. Asked and answered. 09:43AM

11 A. It will depend a great deal on the number of 09:43AM

12 other factors. 09:43AM

13 Q. Okay. Do you have your CV in front of you? 09:43AM

14 A. I don't. 09:43AM

15 Q. Okay. Let me give it to you. This has 09:43AM

16 already been marked as Exhibit 1034 in this case. 09:43AM

17 Is this an accurate CV, Dr. Larrea? 09:44AM

18 A. Yes, I believe it is. 09:44AM

19 Q. How many peer-reviewed publications do you 09:44AM

20 have? 09:44AM

21 A. I have five. 09:44AM

22 Q. Do any of them relate to PacBio Sequencing? 09:44AM

23 A. None of them relate to PacBio Sequencing? 09:44AM

24 Q. Have you ever been faculty at any college or 09:44AM

25 university? 09:44AM

1 A. I have never been faculty at any college or 09:44AM
2 university. 09:45AM
3 Q. Have you ever served as an expert witness 09:45AM
4 before? 09:45AM
5 A. I have never served as an expert witness 09:45AM
6 before. 09:45AM
7 Q. Do you have any patents or patent applications 09:45AM
8 you've filed as an inventor? 09:45AM
9 A. I do not have any patents. 09:45AM
10 Q. You are a trainer for PacBio. Is that 09:45AM
11 correct? 09:45AM
12 A. I was initially a trainer at PacBio, that is 09:45AM
13 correct. 09:45AM
14 Q. And now you are a paid professional 09:45AM
15 consultant. Is that right? 09:45AM
16 A. Consulting is one of the things that I 09:45AM
17 currently do, that is correct. 09:45AM
18 Q. What else do you do? 09:45AM
19 A. I also work professionally. I work at Galatea 09:45AM
20 Bio. 09:46AM
21 Q. What is Galatea Bio? 09:46AM
22 A. It is a company building a biobank located in 09:46AM
23 Miami, Florida. 09:46AM
24 Q. What is a biobank? 09:46AM
25 A. It is a large store of samples. 09:46AM

1 Q. Do you do any sequencing work as part of your 09:46AM
2 role at Galatea Bio? 09:46AM
3 A. We do quite a bit of sequencing work. 09:46AM
4 Q. Do you do that in-house? 09:46AM
5 A. We do some sequencing in-house. 09:46AM
6 Q. What platforms do you use at Galatea Bio? 09:46AM
7 A. We are currently using the Illumina platform. 09:46AM
8 Q. Have you ever used the Pacific Biosciences 09:46AM
9 platform at Galatea Bio? 09:46AM
10 A. We have not used the Pacific Biosciences 09:46AM
11 platform at Galatea Bio. 09:46AM
12 Q. Your CV says you have experience in PacBio and 09:46AM
13 Illumina sequencing? 09:46AM
14 A. I do. 09:46AM
15 Q. Okay. Can you describe your familiarity with 09:46AM
16 Illumina sequencing? 09:47AM
17 A. One of the things I've done is designed 09:47AM
18 experiments for Illumina sequencing. 09:47AM
19 Q. Are you an expert in Illumina sequencing? 09:47AM
20 A. I would consider myself knowledgeable in 09:47AM
21 Illumina sequencing. 09:47AM
22 Q. But not an expert? 09:47AM
23 A. Is there a definition of expert? 09:47AM
24 Q. What's your definition? I'm really just 09:47AM
25 asking you if you would consider yourself an expert in 09:48AM

1 Illumina sequencing. 09:48AM

2 A. I would say I know more than most about 09:48AM

3 Illumina sequencing. 09:48AM

4 Q. More than most people? 09:48AM

5 A. Um-hmm. 09:48AM

6 Q. Did you ever innovate any new hardware for 09:48AM

7 Pacific Biosciences? 09:48AM

8 A. No, I did not. 09:48AM

9 Q. Did you ever innovate any new sequencing 09:48AM

10 methods for Pacific Biosciences? 09:48AM

11 MR. MILLS: I'm going to issue a 09:48AM

12 cautionary instruction to the witness. 09:48AM

13 To the extent that your answers to questions 09:49AM

14 about your work at Pacific Biosciences are controlled 09:49AM

15 by any nondisclosure agreement, I caution you to abide 09:49AM

16 by the terms of any such agreement if one exists. 09:49AM

17 Q. (By Mr. Powers) The question was: Did you 09:49AM

18 ever innovate any new sequencing methods for Pacific 09:49AM

19 Biosciences? 09:49AM

20 A. I was part of the team developing new library 09:49AM

21 protocols at Pacific Biosciences. 09:49AM

22 Q. But you're not on any patent application or 09:49AM

23 patent that issued off of that work. Correct? 09:49AM

24 A. That is correct. 09:50AM

25 Q. How many people were working on that team that 09:50AM

1 was developing new library protocols at Pacific 09:50AM
2 Biosciences? 09:50AM
3 A. One of them had 10. 09:50AM
4 Q. Was there another team? 09:50AM
5 A. Yes. 09:50AM
6 Q. Okay, how many people were on the other team? 09:50AM
7 A. I can't recall an exact number. I'd say about 09:50AM
8 15. 09:50AM
9 Q. Okay, so you worked on teams at Pacific 09:50AM
10 Biosciences that comprised between 10 and 15 people. 09:50AM
11 Is that right? 09:50AM
12 A. In those projects, yes. 09:50AM
13 Q. Do you have in front of you in your binder, 09:50AM
14 there, the Travers documents? Travers 2010 and Travers 09:51AM
15 075? 09:51AM
16 A. Uh-huh. 09:51AM
17 Q. I'd like to turn you to Travers 2010. And 09:51AM
18 that is Exhibit 1021. And I'd like to direct you to 09:51AM
19 page four of the document. On the right side of that 09:51AM
20 page there's a paragraph that starts out: 09:51AM
21 The circular molecules were then applied to 09:52AM
22 SMRT sequencing described in (1). In brief, this 09:52AM
23 system utilizes a number of recently developed 09:52AM
24 technologies to enable multiplex single-molecule 09:52AM
25 observation. 09:52AM

1 Do you see that language? 09:52AM

2 A. I do. 09:52AM

3 Q. So SMRT sequencing was nascent, cutting-edge 09:52AM

4 technology in 2010. Correct? 09:52AM

5 MR. MILLS: Objection to form. Scope. 09:52AM

6 A. Could you repeat the way you described it? 09:52AM

7 Q. Sure. This passage says that these 09:52AM

8 technologies were recently developed. 09:52AM

9 So my question was: SMRT sequencing was 09:52AM

10 nascent, cutting-edge technology in 2010. Correct? 09:52AM

11 MR. MILLS: Same objections. 09:53AM

12 A. In 2010 PacBio was cutting-edge technology. 09:53AM

13 Q. So you agree that it was cutting-edge 09:53AM

14 technology as of 2010? 09:53AM

15 A. Yes. 09:53AM

16 Q. I would like to hand you -- first I'd like to 09:53AM

17 direct you to Travers 2010, Figure 1. Figure 1A. 09:53AM

18 Do you see that? 09:53AM

19 A. I do. 09:53AM

20 Q. Is that a SMRTbell template? 09:53AM

21 A. Yes, that is a SMRTbell template. 09:54AM

22 Q. I'd like to hand you just a blowup of that 09:54AM

23 figure, and we'll mark this as Larrea Exhibit 2. 09:54AM

24 (Exhibit 2 was marked for identification.) 09:54AM

25 Q. (BY MR. POWERS) Are you familiar with this 09:54AM

1 SMRTbell structure? 09:54AM

2 A. Yes, I am. 09:54AM

3 Q. Do you agree that a SMRTbell is topologically 09:54AM

4 circular? 09:54AM

5 A. Yes. The SMRTbell drawing here is 09:54AM

6 topologically circular. 09:54AM

7 Q. But it's a linear molecule. Correct? 09:55AM

8 A. SMRTbell is not a linear molecule. 09:55AM

9 Q. Do you agree that it resembles a linear 09:55AM

10 molecule? 09:55AM

11 MR. MILLS: Objection to scope. 09:55AM

12 A. It does not resemble a linear molecule. 09:55AM

13 Q. So taking the exhibit put in front of you, can 09:55AM

14 you circle the primer/polymerase binding sites on the 09:56AM

15 SMRTbell? And you can use my pen. 09:56AM

16 A. Sorry, can you repeat the question? 09:56AM

17 Q. Yes. Just circle the area where the 09:56AM

18 primer/polymerase binding site is on the SMRTbell. 09:56AM

19 MR. MILLS: Objection to form. 09:56AM

20 A. (Indicating). 09:56AM

21 Q. (By Mr. Powers) Do you agree that polymerase 09:56AM

22 binding to a template is a stochastic process? 09:56AM

23 MR. MILLS: Objection to form and scope. 09:56AM

24 A. (Pause.) 09:56AM

25 Q. (By Mr. Powers) I'll withdraw the question. 09:57AM

1 Dr. Larrea, I'm going to hand you what we will 09:57AM
2 mark Larrea Exhibit 3. 09:57AM
3 (Exhibit 3 was marked for identification.) 09:57AM
4 Q. (By Mr. Powers) This is a schematic of a 09:57AM
5 linear single-stranded RCR product generated from a 09:57AM
6 polymerase going around a SMRTbell three times. 09:57AM
7 Are you with me? 09:58AM
8 A. I am. 09:58AM
9 Q. Can you please circle all those 09:58AM
10 primer/polymerase binding sites? 09:58AM
11 A. Sorry, can you repeat what it is that I'm 09:58AM
12 looking at? 09:58AM
13 Q. So what you're looking at is a schematic of a 09:58AM
14 linear single-stranded RCR product generated from a 09:58AM
15 polymerase going around three times a SMRTbell. 09:58AM
16 So I'd like you to circle all the 09:58AM
17 primer/polymerase binding sites on this product. 09:58AM
18 A. All of the potential primer/polymerase binding 09:59AM
19 sites? 09:59AM
20 Q. Yes, please. 09:59AM
21 A. Is there a legend that goes with the figure? 09:59AM
22 Q. Well, let's use the same colors that we have 09:59AM
23 here in the Travers SMRTbell. 09:59AM
24 A. So should that also be on here? 09:59AM
25 Q. It doesn't have to be. I can represent to you 09:59AM

1 that the green portions represent the hairpin-linkers. 09:59AM
2 And the color-coded represent the strands. 09:59AM
3 A. (Indicating). 10:00AM
4 Q. Thank you. And I'd like to hand you what we 10:00AM
5 will mark Larrea Exhibit 4. 10:00AM
6 (Exhibit 4 was marked for identification.) 10:00AM
7 Q. (By Mr. Powers) It's the same idea. It's a 10:00AM
8 schematic of a linear single-stranded RCR product 10:00AM
9 generated from a polymerase going four times around the 10:00AM
10 SMRTbell. Sorry, strike that. 10:00AM
11 It's five times around the SMRTbell. So I'd 10:00AM
12 like you to mark all of the polymerase/primer binding 10:00AM
13 sites on this product. 10:01AM
14 A. And once again, the color scheme is the same? 10:01AM
15 Q. Yes, sir. 10:01AM
16 A. (Indicating). 10:01AM
17 Q. Okay. And so those arrows that you've marked 10:01AM
18 indicate the polymerase/primer binding sites on 10:01AM
19 Exhibits 3 and 4. Correct? 10:01AM
20 A. They represent, according to the legend that I 10:01AM
21 was given, putative for potential annealing binding 10:01AM
22 sites. 10:01AM
23 Q. And I'd like to hand you what we will mark 10:01AM
24 Larrea Exhibit 5. 10:01AM
25 (Exhibit 5 was marked for identification.) 10:02AM

1	Q. (By Mr. Powers) This is a schematic of a	10:02AM
2	linear single-stranded RCR product generated from a	10:02AM
3	polymerase going ten times around a SMRTbell.	10:02AM
4	So could you please mark all the	10:02AM
5	polymerase/primer binding sites on this product.	10:02AM
6	A. Can I just mark it with a line instead of an	10:02AM
7	arrow?	10:02AM
8	Q. Sure.	10:02AM
9	A. Okay.	10:02AM
10	Q. So if you look at Exhibits 3, 4, and 5,	10:02AM
11	polymerase binding on each of these structures would be	10:03AM
12	stochastic. Correct?	10:03AM
13	A. Yes.	10:03AM
14	Q. And there would be no way to control which of	10:03AM
15	the binding sites bound polymerase over any other site.	10:03AM
16	Correct?	10:03AM
17	MR. MILLS: Objection to form. Scope.	10:03AM
18	A. Based on this cartoon, that is correct.	10:03AM
19	Q. (By Mr. Powers) I'd like to take you to	10:03AM
20	Figure 3 on Travers. You can put those to the side.	10:03AM
21	Are you there, sir?	10:04AM
22	A. Yes.	10:04AM
23	Q. Figure 3 shows the product sizes produced by	10:04AM
24	an RCR reaction using a SMRTbell template. Correct?	10:04AM
25	A. Yes, that is correct.	10:04AM

1 Q. Travers 2010 doesn't disclose using the 10:04AM
2 products from Figure 3 in any sequencing reaction, 10:04AM
3 right? 10:04AM
4 MR. MILLS: Objection to form. Scope. 10:04AM
5 Q. (By Mr. Powers) To be clear, Travers 2010 10:05AM
6 doesn't disclose using the products from Figure 3 in 10:05AM
7 any sequencing reaction? 10:05AM
8 MR. MILLS: Same objection. 10:05AM
9 A. The products from Figure 3 are not then 10:05AM
10 sequenced in this paper because the data is not 10:05AM
11 disclosed. 10:05AM
12 Q. (By Mr. Powers) Do you know what a Poisson 10:05AM
13 distribution is? 10:05AM
14 A. I do. 10:05AM
15 Q. What is it? 10:05AM
16 A. It is a mathematical function and it is the 10:05AM
17 basis on which PacBio loading is done. 10:05AM
18 Q. Do you agree that the CMW loading as described 10:05AM
19 in Travers 2010 and Travers 075 follow a Poisson 10:05AM
20 distribution? 10:06AM
21 A. I agree that that was the goal of the loading. 10:06AM
22 Q. And the goal of a Poisson loading in the 10:06AM
23 PacBio system is to obtain as many zero-mode waveguides 10:06AM
24 as possible loaded with a single SMRTbell template. 10:06AM
25 Correct? 10:06AM

1 A. It was the method used to maximize single 10:06AM
2 loading -- 10:06AM
3 Q. When you say -- 10:06AM
4 A. -- of SMWs. 10:07AM
5 Q. I'm sorry. When you say "single loading," you 10:07AM
6 mean a single polymerase-bound template in a zero-mode 10:07AM
7 waveguide. Right? 10:07AM
8 A. In a zero-mode waveguide. That's correct. 10:07AM
9 MR. POWERS: Let's take a break. We've 10:07AM
10 been going for about an hour. 10:07AM
11 (Recess taken 10:07 a.m. to 10:24 a.m.) 10:07AM
12 Q. (By Mr. Powers) Dr. Larrea, did you discuss 10:24AM
13 the substance of your testimony with counsel during the 10:24AM
14 break? 10:24AM
15 A. We did not. 10:24AM
16 Q. Did counsel discuss the manner of your 10:24AM
17 testimony with you during the break? 10:24AM
18 A. They did not. 10:24AM
19 Q. Let's continue on with Travers 2010. Let me 10:24AM
20 direct you to page 2 of Travers 2010. 10:24AM
21 Under the heading on the right side, it says 10:24AM
22 "Sequencing Reactions." 10:24AM
23 Travers 2010 uses biotinylated polymerases. 10:25AM
24 Correct? 10:25AM
25 A. Yes, that is correct. 10:25AM

1 Q. And on page 4 of Travers 2010 it goes on to 10:25AM
2 say -- in the right column, the first paragraph says 10:25AM
3 "The circular molecules." 10:25AM
4 If you continue reading down through that 10:25AM
5 paragraph, you'll agree that Travers 2010 uses 10:25AM
6 streptavidin coating on its zero-mode waveguides. 10:26AM
7 Correct? 10:26AM
8 A. Yes. That's a streptavidin biotin. 10:26AM
9 (Court Reporter requested the witness speak 10:26AM
10 louder.) 10:26AM
11 A. That's a streptavidin biotin interaction. 10:26AM
12 Q. (By Mr. Powers) So the polymerases are bound 10:26AM
13 to the bottom of a zero-mode waveguide by a 10:26AM
14 streptavidin biotin interaction. Correct? 10:26AM
15 MR. MILLS: Objection, form. 10:26AM
16 A. At the time, I believe that was the correct 10:26AM
17 interaction. 10:26AM
18 Q. (By Mr. Powers) Does the SMRT sequencing 10:26AM
19 described in Travers 2010 use a flow cell? 10:27AM
20 A. Travers 2010 does not use a flow cell. 10:27AM
21 Q. Do SMRTbell templates have flow cell binding 10:27AM
22 sites? 10:27AM
23 MR. MILLS: Objection to form. 10:27AM
24 A. (Pause.) The SMRTbell template doesn't 10:27AM
25 really -- I guess I don't understand the question 10:28AM

1 biotin-streptavidin interaction. 10:30AM

2 Q. Do you agree that the interaction between 10:31AM

3 streptavidin and biotin is one of the strongest 10:31AM

4 noncovalent interactions in nature? 10:31AM

5 A. The strong interaction between biotin and 10:31AM

6 streptavidin is often used for a variety of reasons. 10:31AM

7 Whether it's the strongest, I can't speak to. 10:31AM

8 Q. Do you agree that it's very strong? 10:31AM

9 A. It's a strong interaction. 10:31AM

10 Q. Are you familiar with the term "polymerase 10:31AM

11 read" as used in PacBio sequencing? 10:32AM

12 MR. MILLS: Objection to form. 10:32AM

13 A. Yes, I am familiar with the term "polymerase 10:32AM

14 read." 10:32AM

15 Q. (By Mr. Powers) Do you agree that a 10:32AM

16 polymerase read is a sequence of nucleotides 10:32AM

17 incorporated by the DNA polymerase while reading a 10:32AM

18 template such as a circular SMRTbell template? 10:32AM

19 MR. MILLS: Objection to scope. 10:32AM

20 A. Sorry, that's a long sentence. 10:32AM

21 Can you repeat -- 10:32AM

22 Q. (By Mr. Powers) Sure. 10:32AM

23 A. -- to make sure that all parts of that make 10:32AM

24 sense? 10:32AM

25 Q. Happy to. 10:32AM

1 Do you agree that a polymerase read is a 10:32AM
2 sequence of nucleotides incorporated by the DNA 10:32AM
3 polymerase while reading a template such as a circular 10:32AM
4 SMRTbell template? 10:33AM
5 MR. MILLS: Objection. Scope. 10:33AM
6 A. A polymerase read is the series of nucleotides 10:33AM
7 incorporated -- well, by a polymerase bound to a primer 10:33AM
8 template in a ZMW specifically for PacBio data. 10:33AM
9 Q. (By Mr. Powers) Do you agree that polymerase 10:33AM
10 reads can include sequences from adaptors and from one 10:33AM
11 or multiple passes around the circular template? 10:33AM
12 MR. MILLS: Objection to scope. 10:34AM
13 A. A read from a polymerase -- sorry, can you 10:34AM
14 repeat the phrase once more? 10:34AM
15 Q. (By Mr. Powers) Sure. 10:34AM
16 Do you agree that a polymerase read can 10:34AM
17 include sequences from adapters and from one or 10:34AM
18 multiple passes around a circular template? 10:34AM
19 MR. MILLS: Objection to scope. 10:34AM
20 A. In PacBio sequencing, a read is the 10:34AM
21 nucleotides that come from an individual polymerase 10:35AM
22 going through the incorporation events. So it may 10:35AM
23 have, yeah, the parts that you describe. 10:35AM
24 Q. (By Mr. Powers) Are you familiar with the 10:35AM
25 term "polymerase read length" as used in PacBio 10:35AM

1 sequencing? 10:35AM

2 A. I am. 10:35AM

3 Q. Do you agree that polymerase read length is 10:35AM

4 the total number of bases produced from a zero-mode 10:35AM

5 waveguide after trimming the low-quality regions? 10:36AM

6 MR. MILLS: Objection to scope. 10:36AM

7 A. Yes. Polymerase read length is the set of 10:36AM

8 nucleotides once the sort of low-quality regions are 10:37AM

9 removed. 10:37AM

10 Q. (By Mr. Powers) Are you familiar with the term 10:37AM

11 "subread" in PacBio sequencing? 10:37AM

12 MR. MILLS: Objection to scope. 10:37AM

13 A. I am familiar with the term "subread." 10:37AM

14 Q. (By Mr. Powers) Do you agree that in the 10:37AM

15 PacBio workflow a polymerase read is partitioned into 10:37AM

16 subreads? 10:37AM

17 MR. MILLS: Objection. Scope. 10:37AM

18 A. In PacBio it may be, but it would depend on 10:37AM

19 the exact experiment. 10:38AM

20 Q. (By Mr. Powers) I'm going to hand you what we 10:38AM

21 can mark Larrea Exhibit 6. 10:38AM

22 (Exhibit 6 was marked for identification.) 10:38AM

23 Q. (By Mr. Powers) So if we go to page 2. 10:38AM

24 MR. MILLS: Before you ask questions, we 10:38AM

25 just need a moment to review the document. 10:38AM

1 MR. POWERS: Sure. 10:38AM

2 MR. MILLS: So I'll object to the document 10:39AM

3 on the basis of hearsay. 10:39AM

4 MR. POWERS: Okay. 10:39AM

5 Q. (By Mr. Powers) Are you familiar with the 10:39AM

6 terms P0, P1 and P2 in the context of Pacific 10:40AM

7 Biosciences sequencing? 10:40AM

8 A. Yes, I am. 10:40AM

9 Q. Do you agree that P1 means that there is a 10:40AM

10 polymerase read from that zero-mode waveguide? 10:40AM

11 MR. MILLS: Objection to scope. 10:40AM

12 Q. (By Mr. Powers) I'll direct you to 10:40AM

13 productivity definition on page 2 of Exhibit 6. 10:40AM

14 MR. MILLS: Same objection. 10:40AM

15 A. Yes, P1 is that there is a polymerase read 10:40AM

16 generated by that ZMW. 10:41AM

17 Q. (By Mr. Powers) And P0 means that a zero-mode 10:41AM

18 waveguide did not produce a read and is presumed to be 10:41AM

19 lacking a polymerase. Correct? 10:41AM

20 MR. MILLS: Objection to scope. 10:41AM

21 A. P0 did not include -- did not generate a read 10:41AM

22 in the ZMW. 10:41AM

23 Q. (By Mr. Powers) And that's because it's 10:41AM

24 presumed to be lacking a polymerase, right? 10:41AM

25 MR. MILLS: Objection. Scope. 10:41AM

1 A. Yes, the assumption is that there is no 10:42AM
2 polymerase in that ZMW. 10:42AM

3 Q. (By Mr. Powers) In P=2 means that -- it means 10:42AM
4 other, and that the signal collected from the zero-mode 10:42AM
5 waveguide was not conducive to efficient base-calling, 10:42AM
6 possibly due to multiple template-polymerase complexes 10:42AM
7 bound in the ZMW, high background signal, and so on. 10:42AM

8 Correct? 10:42AM

9 MR. MILLS: Objection. Scope. 10:42AM

10 A. P2 was often used as a broad bucket for things 10:42AM
11 that were neither P0 or P1. 10:43AM

12 Q. (By Mr. Powers) And one of the reasons 10:43AM
13 something could be a P2 is because there were multiple 10:43AM
14 template-polymerase complexes bound in the ZMW. Right? 10:43AM

15 MR. MILLS: Objection to scope. 10:43AM

16 A. One of the reasons that it could be P2 would 10:43AM
17 be that it just meant that no usable data was generated 10:44AM
18 from that ZMW. 10:44AM

19 Q. (By Mr. Powers) Look further down on the page 10:44AM
20 where it says "subreads." Are you there? 10:44AM

21 A. Um-hmm. 10:44AM

22 Q. Do you agree that in the PacBio workflow 10:44AM
23 polymerase read is partitioned into subreads? 10:44AM

24 MR. MILLS: Objection to form. Scope. 10:44AM

25 A. Is there something in my declaration that 10:44AM

1 specifically talks about subreads and reads and 10:44AM
2 polymerase read lengths? Or am I missing something? 10:44AM
3 Q. (By Mr. Powers) Well, you're supposed to be 10:45AM
4 an expert in PacBio sequencing. Right? 10:45AM
5 A. That is correct. 10:45AM
6 Q. And this is PacBio's glossary of terms. 10:45AM
7 So subreads says: Each polymerase read is 10:45AM
8 partitioned to form one or more subreads which contain 10:45AM
9 sequence from a single pass of a polymerase on a single 10:45AM
10 strand of an insert within a SMRTbell template and no 10:45AM
11 adapter sequences. 10:45AM
12 Did I read that correctly? 10:45AM
13 MR. MILLS: Objection to form. Outside 10:45AM
14 the scope. 10:45AM
15 A. You read that correctly. 10:45AM
16 Q. Do you disagree with Pacific Bioscience's 10:45AM
17 definition of subreads here? 10:45AM
18 MR. MILLS: Objection to form and scope. 10:45AM
19 A. I agree that that is the definition given 10:45AM
20 here. 10:46AM
21 Q. (By Mr. Powers) Okay. And you don't disagree 10:46AM
22 with that definition, do you? 10:46AM
23 MR. MILLS: Objection to form. Scope. 10:46AM
24 A. Subreads were a convenient way of describing 10:46AM
25 portions of the read. 10:47AM

1 Q. (By Mr. Powers) Let's go to Exhibit 1035. Is 10:47AM
2 that one of the ones that you have in your binder? 10:47AM
3 A. Yes. 10:47AM
4 Q. Let's go to page 3 of the document. Down 10:47AM
5 towards the bottom of the text on page 3 it says: The 10:48AM
6 hairpin dimers formed during this process are removed 10:48AM
7 at the end of the protocol using a magnetic bead 10:48AM
8 purification step with size-selective conditions. 10:48AM
9 Did I read that correctly? 10:48AM
10 A. You did. 10:48AM
11 Q. So Exhibit 1035 discloses using magnetic beads 10:48AM
12 to remove hairpin dimers. Correct? 10:48AM
13 A. That is correct. 10:48AM
14 Q. And a hairpin dimer is comprised of two 10:48AM
15 hairpins? 10:48AM
16 A. With no intervening insert. 10:48AM
17 Q. It's just two hairpins attached together? 10:48AM
18 A. Attached together. 10:48AM
19 Q. Let's go to page 34 of this document. I'm 10:48AM
20 sorry, 24, rather. Page 24. 10:48AM
21 There's a section there that says: AMPure PB 10:49AM
22 beads size selection and purification. 10:49AM
23 Do you see that? 10:49AM
24 A. Yes. 10:49AM
25 Q. Do you know what AMPure PB beads are? 10:49AM

1	A. Yes.	10:49AM
2	Q. Does the PB stand for Pacific Biosciences?	10:49AM
3	A. It does.	10:49AM
4	Q. Do you know if AMPure PB beads were made	10:49AM
5	specifically for Pacific Biosciences?	10:49AM
6	A. AMPure PB beads were made specifically for	10:49AM
7	PacBio.	10:49AM
8	Q. In page 24 of Exhibit 1035 discloses using	10:49AM
9	AMPure PB beads for purifying SMRTbell templates.	10:50AM
10	Correct?	10:50AM
11	A. Sorry, can you repeat that?	10:50AM
12	Q. Yeah.	10:50AM
13	A. In the moment, I was just zoning out.	10:50AM
14	Q. No problem. It happens to the best of us.	10:50AM
15	Page 24 of Exhibit 1035 discloses using AMPure	10:50AM
16	PB beads for purifying SMRTbell templates. Correct?	10:50AM
17	A. Yes.	10:50AM
18	Q. If you look down on page 24 --	10:50AM
19	A. One point of clarification. The AMPure PB	10:50AM
20	beads were a derivative product off the AMPure beads,	10:50AM
21	and you could make AMPure PB beads by washing the	10:50AM
22	AMPure beads.	10:50AM
23	Q. Interesting.	10:50AM
24	A. And there was a protocol out for that.	10:51AM
25	Q. Okay. At the top of page 24, rather, there's	10:51AM

Page 44

1 a -- the first paragraph. And a little bit more than 10:51AM
2 halfway through that paragraph it states: Perform all 10:51AM
3 purification steps at room temperature to adequately 10:51AM
4 remove enzymes, exonucleases, ligases, et cetera, and 10:51AM
5 ligation products smaller than 0.4KB (EG adapter 10:51AM
6 dimers). 10:51AM
7 Did I read that correctly? 10:51AM
8 A. Yes. 10:51AM
9 Q. So AMPure PB beads remove ligation products 10:51AM
10 smaller than 0.4KB. Correct? 10:51AM
11 A. They may. It depends on the amount of AMPure 10:51AM
12 beads that you use. 10:52AM
13 Q. In this protocol it's for 0.4KB and smaller. 10:52AM
14 Correct? 10:52AM
15 A. Correct. 10:52AM
16 Q. Do you have Exhibit 1036 in your binder? 10:52AM
17 A. Yes, I do. 10:52AM
18 Q. It's the supplemental, there's all this data 10:52AM
19 in the back. 10:52AM
20 Do you have a copy? 10:52AM
21 A. I removed the supplement. Unless you -- you 10:52AM
22 don't need me to have the supplement. 10:52AM
23 Q. I'm not going to ask you questions about the 10:52AM
24 supplement. I'll ask you questions about page 3. 10:52AM
25 Page 3 on the bottom right, it says: We constructed 10:53AM

1 DNA libraries comprising SMRTbell constructs, each of 10:53AM
2 which was bound to a DNA polymerase and sequenced in a 10:53AM
3 manner similar to that described previously using the 10:53AM
4 PacBio RS sequencing system. 10:53AM
5 Did I read that correctly? 10:53AM
6 A. You did. 10:53AM
7 Q. So these researchers used a SMRTbell template 10:53AM
8 format. Correct? 10:53AM
9 A. They did. 10:53AM
10 Q. Let's go to page 12. So I guess I am going to 10:53AM
11 ask you some questions about the supplemental, just not 10:54AM
12 the voluminous data in the very back. 10:54AM
13 The second paragraph on that page, page 12, 10:54AM
14 starts off with the word "approximately." 10:54AM
15 Do you see that? 10:54AM
16 A. I do. 10:54AM
17 Q. All right. The next sentence in that 10:54AM
18 paragraph says: A SMRTbell is a double-stranded DNA 10:54AM
19 template capped by single-stranded hairpin loops at 10:54AM
20 both ends. 10:54AM
21 And cites to Travers 2010. Correct? 10:54AM
22 A. It does. 10:55AM
23 Q. The language continues: Genomic DNA was 10:55AM
24 mechanically sheared to a target size of 2,000 base 10:55AM
25 pairs using a Covaris S2 system according to the 10:55AM

1 manufacturer's instructions. 10:55AM

2 Correct? 10:55AM

3 A. Correct. 10:55AM

4 Q. Does this mean that the sheared fragments 10:55AM

5 generated are all approximately the same size? 10:55AM

6 MR. MILLS: Objection to scope. 10:55AM

7 A. When mechanically shearing, you'll get a 10:55AM

8 distribution of sizes. 10:55AM

9 (Reporter requested clarification and for the 10:55AM

10 witness to speak louder.) 10:55AM

11 Q. (By Mr. Powers) Indeed you will. And if you 10:55AM

12 go to page 13, there's a sentence that says: The means 10:56AM

13 from the fragment distributions were between 17,000 and 10:56AM

14 2,000 BP. 10:56AM

15 Correct? 10:56AM

16 MR. MILLS: Objection to scope. 10:56AM

17 A. That is what it says there. 10:56AM

18 Q. (By Mr. Powers) If you go back to page 12, at 10:56AM

19 the very bottom, there, three lines up, it says: 10:57AM

20 Sheared samples were then concentrated and 10:57AM

21 size-selected, i.e., fragments greater than 100 base 10:57AM

22 pairs were retained by solid-phase reversible 10:57AM

23 immobilization SPRI paramagnetic bead-based technology. 10:57AM

24 And cites to AMPure XP. Correct? 10:58AM

25 A. Correct. 10:58AM

1 Q. Do you know what AMPure XP Beads are? 10:58AM

2 A. Yes. 10:58AM

3 Q. And, here, Chin discloses that using AMPure XP 10:58AM

4 beads fragments greater than 100 base pairs are 10:58AM

5 retained. 10:58AM

6 Correct? 10:58AM

7 A. Under the conditions that they were using 10:58AM

8 them, that's what they did. 10:58AM

9 MR. POWERS: Would you like to take a 10:58AM

10 quick break? 10:58AM

11 THE WITNESS: I just need coffee. 10:58AM

12 MR. POWERS: Go ahead. We can go off the 10:58AM

13 record. 10:58AM

14 (Off the record.) 10:59AM

15 Q. (By Mr. Powers) Do you have Travers 075 in 10:59AM

16 your binder? 11:00AM

17 A. I do. 11:00AM

18 Q. Great. Okay. So you have in front of you 11:00AM

19 Travers 075. That's Exhibit 1018; right? 11:00AM

20 A. I do. 11:00AM

21 Q. Go to paragraph 145 in Travers. It says: The 11:00AM

22 ligation products were concentrated by ethanol 11:00AM

23 precipitation and then applied to a CHROMA SPIN-1000 11:00AM

24 column to remove any templates that contained no 11:00AM

25 inserts or short inserts. 11:00AM

1 Do you see that? 11:00AM

2 A. I do. 11:00AM

3 Q. Do you know what CHROMA SPIN-1000 columns are? 11:00AM

4 A. Yes. 11:00AM

5 Q. Travers discloses using CHROMA SPIN-1000 11:00AM

6 columns to remove templates with no insert or short 11:00AM

7 inserts. Correct? 11:01AM

8 A. Yes. 11:01AM

9 Q. Do you know who the manufacturer of CHROMA 11:01AM

10 SPIN-1000 columns is? 11:01AM

11 A. I don't recall. 11:01AM

12 Q. I'd like to provide you with what we'll mark 11:01AM

13 Larrea Exhibit 7. 11:01AM

14 (Exhibit 7 was marked for identification.) 11:01AM

15 Q. (By Mr. Powers) Will you go to -- 11:01AM

16 MR. MILLS: Before you ask questions, give 11:01AM

17 me just a moment. 11:01AM

18 MR. POWERS: Sure. 11:01AM

19 MR. MILLS: I object to the document on 11:02AM

20 the basis of authenticity, foundation, and hearsay. 11:02AM

21 MR. POWERS: So noted. 11:02AM

22 Q. (By Mr. Powers) Have you seen this document 11:02AM

23 before, Dr. Larrea? 11:02AM

24 A. I can't recall. 11:02AM

25 Q. If you look on the first page of the document, 11:02AM

1 it says it was published '09 September 2004. Correct? 11:02AM

2 MR. MILLS: Objection. Foundation. 11:02AM

3 A. That is what it says. 11:02AM

4 Q. (By Mr. Powers) If you look at the bottom of 11:02AM

5 page 2, it says "Clone Tech Laboratories, Inc." 11:02AM

6 Correct? 11:02AM

7 MR. MILLS: Same objection. 11:03AM

8 A. That is what it says. 11:03AM

9 Q. (By Mr. Powers) Are you familiar with Clone 11:03AM

10 Tech Laboratories? 11:03AM

11 A. I have worked with their products in the past. 11:03AM

12 Q. Does this document refresh your recollection 11:03AM

13 about the manufacturer of CHROMA SPIN columns? 11:03AM

14 A. It does in that it says it. I didn't remember 11:03AM

15 it. 11:03AM

16 Q. But now you do? 11:03AM

17 A. Sure. 11:03AM

18 Q. If you go to page 6 of the document, for 11:03AM

19 CHROMA SPIN-1000 columns it says -- the document says: 11:04AM

20 To use CHROMA SPIN-1000 for purifying DNA molecules 11:04AM

21 greater than 1,350 base pairs. Correct? 11:04AM

22 A. That is what it says there. 11:04AM

23 Q. And it says to use CHROMA SPIN-1000 for 11:04AM

24 removing RNA and DNA molecules less than 300 base 11:04AM

25 pairs. Correct? 11:04AM

1 Q. (By Mr. Powers) If you go to page 2 -- 11:07AM
2 MR. MILLS: Sorry. Just a moment. 11:07AM
3 Q. (By Mr. Powers) The second paragraph starts 11:07AM
4 out with slab-gel electrophoresis. 11:07AM
5 MR. MILLS: Sorry. Just a moment. 11:07AM
6 I object to the document on the basis of 11:07AM
7 hearsay and outside the scope. 11:07AM
8 Q. (By Mr. Powers) Do you see in that paragraph, 11:07AM
9 Doctor, the language: However, typical SGE experiment 11:08AM
10 procedures are quite tedious, which involves pairing 11:08AM
11 the agarose gel 30 to 50 minutes; sample loading one to 11:08AM
12 five minutes; running electrophoresis 30 to 60 minutes; 11:08AM
13 gel staining 15 to 30 minutes; and imaging DNA bands, 11:08AM
14 one to five minutes. 11:08AM
15 Did I read that correctly? 11:08AM
16 MR. MILLS: Objection to form. Scope. 11:08AM
17 Hearsay. 11:08AM
18 A. You did. 11:08AM
19 Q. (By Mr. Powers) The document continues: 11:08AM
20 Consequently, it requires a lot of instruments such as 11:08AM
21 a microwave to dissolve the agarose, 2D gel -- 2D 11:08AM
22 electrophoresis cell and electrophoresis power supply 11:08AM
23 for electrophoresis, gel imaging system to check the 11:08AM
24 DNA bands in the gel, and hence, traditional SGE is 11:09AM
25 labor intensive, time-consuming, skill-dependent, and 11:09AM

1 poor sensitivity. 11:09AM

2 Did I read that correctly? 11:09AM

3 A. You did. 11:09AM

4 MR. MILLS: Objection to form. Scope. 11:09AM

5 Hearsay. 11:09AM

6 Q. (By Mr. Powers) The document continues: 11:09AM

7 Furthermore, the DNA molecules are usually visualized 11:09AM

8 under ultraviolet light by staining with appropriated 11:09AM

9 dyes such as ethidium bromide, in a traditional SGE 11:09AM

10 experiment. But both ultraviolet and ethidium bromide 11:09AM

11 are suspected to be carcinogen. 11:09AM

12 Did I read that correctly? 11:09AM

13 MR. MILLS: Objection to form, scope, 11:09AM

14 hearsay. 11:09AM

15 A. You read that correctly. 11:09AM

16 Q. (By Mr. Powers) Let's take out Travers again. 11:09AM

17 This is Travers 075. 11:10AM

18 MR. POWERS: Actually, can we take a short 11:10AM

19 break? 11:10AM

20 MR. MILLS: Sure. 11:10AM

21 (Recess taken 11:10 a.m. - 11:19 a.m.) 11:10AM

22 Q. (By Mr. Powers) Do you have Travers 075? 11:19AM

23 A. Exhibit 1018? 11:19AM

24 Q. Yes, sir. 11:19AM

25 A. Okay. 11:19AM

1 Q. I'll take you to example 1 of Travers. It 11:19AM
2 starts out paragraph 138. Moving onto 139, Travers 11:19AM
3 discusses template construction. 11:19AM
4 In 139 Travers says: Ligation was performed 11:20AM
5 by incubating digested PCR product with a two-fold 11:20AM
6 excess of each hairpin oligo in the presence of T4 DNA 11:20AM
7 ligase NEB at 23 degrees celsius for one hour. 11:20AM
8 Do you see that? 11:20AM
9 A. I do. 11:20AM
10 Q. This created a SMRTbell template. Correct? 11:20AM
11 A. Yes, this would have been an initial SMRTbell. 11:20AM
12 Q. And the sequencing reaction was carried out in 11:21AM
13 example 1 on that SMRTbell product. Correct? 11:21AM
14 A. After the rest of the cleanup steps, there was 11:21AM
15 a sequencing reaction. 11:21AM
16 Q. The SMRTbell products were generated from a 11:21AM
17 PCR product. Correct? 11:21AM
18 A. In this experiment they were, yes. 11:22AM
19 Q. So all of the SMRTbells in example 1 would 11:22AM
20 have been almost exactly the same length. Correct? If 11:22AM
21 not exactly the same length. 11:22AM
22 A. It would depend on the ligation. 11:22AM
23 Q. Do you agree that they would have been very, 11:22AM
24 very close in length? The SMRTbell products? 11:22AM
25 A. The PCR products would have been, yes. 11:22AM

1 Q. The PCR products that went in to make the 11:22AM
2 SMRTbell product. SMRTbell template. 11:22AM
3 A. Yes. But the resulting library may not have 11:23AM
4 been as narrow as the PCR product. Depending on the 11:23AM
5 ligation conditions. 11:23AM
6 Q. Is that because there could have been PCR 11:23AM
7 products ligated to each other? 11:23AM
8 A. That is correct. 11:23AM
9 Q. Ligating PCR products to each other is an 11:23AM
10 undesirable outcome in SMRTbell preparation. Correct? 11:23AM
11 A. In this experiment, it would have been 11:23AM
12 undesirable. 11:23AM
13 Q. If you go to example 2, do you agree that a 11:23AM
14 SMRTbell template was generated containing 162 11:23AM
15 nucleotides in total? 11:23AM
16 A. Yes. 11:23AM
17 Q. So example 2 uses a SMRTbell template for 11:23AM
18 sequencing. Correct? 11:24AM
19 A. It does. 11:24AM
20 Q. If you go to example 3. In example 3 the 11:24AM
21 E. coli genome was sheared to generate 500 to 1500 base 11:24AM
22 pair-sized fragments. Correct? 11:24AM
23 A. That is correct. 11:24AM
24 Q. And paragraph 145 states: Fragments of sample 11:24AM
25 material that failed to ligate with a hairpin at one 11:24AM

1 end or that contained a nick due to incomplete ligation 11:24AM
2 were removed through the use of Exonuclease III and 11:24AM
3 Exonuclease VII. 11:24AM
4 Correct? 11:24AM
5 A. That is what it says. 11:24AM
6 Q. So do you agree that example 3 in Travers 075 11:24AM
7 is using SMRTbell templates for sequencing? 11:24AM
8 A. I do. 11:25AM
9 Q. If you go to example 4, paragraph 150, Travers 11:25AM
10 075 makes use in example 4 of a PCR product. Correct? 11:25AM
11 A. It does. 11:25AM
12 Q. And that PCR product was ligated to hairpin 11:25AM
13 adapters to make a SMRTbell template. Correct? 11:25AM
14 A. That is correct. 11:25AM
15 Q. Paragraph 150 of Travers 075 states that: The 11:26AM
16 SMRTbell template was passed through a CHROMA SPIN-1000 11:26AM
17 column. 11:26AM
18 Correct? 11:26AM
19 A. That is what it says. 11:26AM
20 Q. If you look at Travers 075, paragraph 141. 11:26AM
21 Are you there? 11:27AM
22 A. I am. 11:27AM
23 Q. Travers 075 incorporates by reference U.S. 11:27AM
24 Patent No. -- Patent Application No. 200090024331. 11:27AM
25 Do you see that? 11:27AM

1	A. I do.	11:27AM
2	Q. And Travers says -- the Travers 075 document	11:27AM
3	says that: Published U.S. Patent Application	11:27AM
4	No. 20090024331 is incorporated by reference in its	11:28AM
5	entirety for all purposes.	11:28AM
6	Do you see that?	11:28AM
7	A. I do.	11:28AM
8	Q. Did you review U.S. Patent Application	11:28AM
9	20090024331 before rendering your opinions in this	11:28AM
10	case?	11:28AM
11	A. I did not. I focused my analysis on the four	11:28AM
12	alleged problems that Dr. Satija raises in his	11:28AM
13	declaration.	11:28AM
14	Q. And you didn't review this application before	11:28AM
15	rendering your opinions. Right?	11:29AM
16	A. I did not.	11:29AM
17	Q. Okay. I'm going to hand you what we will mark	11:29AM
18	Larrea Exhibit 9.	11:29AM
19	(Exhibit 9 was marked for identification.)	11:29AM
20	Q. (By Mr. Powers) I'd like to direct you to	11:29AM
21	paragraph 181.	11:29AM
22	First of all, if you look at the front page,	11:29AM
23	first page, you can see that the assignee is Pacific	11:29AM
24	Biosciences in California. Right?	11:29AM
25	A. Sorry, I'm catching up.	11:29AM

Page 57

1 Q. Sure. 11:29AM

2 A. Repeat, please. 11:29AM

3 Q. If you look at the first page of Exhibit 9, 11:30AM

4 you'll see that the assignee is Pacific Biosciences of 11:30AM

5 California. Right? 11:30AM

6 A. Yes. 11:30AM

7 Q. And now going to paragraph 181. 11:30AM

8 A. Sure. 11:30AM

9 Q. And I want you to confirm before you go to 11:30AM

10 paragraph 181 that this application publication 11:30AM

11 20090024331 is the same cited in Travers that we just 11:30AM

12 looked at. The Travers 075 publication. 11:30AM

13 MR. MILLS: Objection. Foundation. 11:30AM

14 Q. (By Mr. Powers) And incorporated by 11:30AM

15 reference. 11:31AM

16 A. Yes. 11:31AM

17 Q. Now let's go to paragraph 181. And the second 11:31AM

18 sentence of that paragraph, says: Because large 11:31AM

19 numbers of reaction locations are being prepared and 11:31AM

20 monitored, it is expected that some system -- excuse 11:31AM

21 me -- it is expected that in some systems some 11:31AM

22 percentage of reaction locations will not provide 11:31AM

23 useful data. This may occur if no reaction enzyme 11:31AM

24 becomes located in a particular ZMW, if more than one 11:31AM

25 reaction enzyme is located in a ZMW, or if a reaction 11:31AM

1 enzyme is otherwise producing problematic data. 11:31AM

2 Did I read that correctly? 11:31AM

3 A. You did read that correctly. 11:31AM

4 MR. POWERS: Let's take a quick break. I 11:31AM

5 think I'm going to pass the witness, but I just want to 11:32AM

6 confirm. 11:32AM

7 (Off the record.) 11:32AM

8 MR. POWERS: We're done. 11:32AM

9 (Recess taken 11:32 a.m. to 12:18 p.m.) 11:32AM

10 * * * 11:32AM

11 12:18PM

12 REDIRECT EXAMINATION 12:18PM

13 BY MR. MILLS: 12:18PM

14 Q. Dr. Larrea, I have a few questions for you. 12:18PM

15 Do you recall being asked questions during 12:18PM

16 your deposition about gel electrophoresis? 12:18PM

17 A. Yes. 12:18PM

18 Q. Is it difficult to perform a gel 12:18PM

19 electrophoresis? 12:18PM

20 MR. POWERS: Objection, leading. 12:18PM

21 A. No. In fact, my wife is currently teaching a 12:18PM

22 high school genetics class in which she's teaching 12:19PM

23 sophomores and juniors how to perform it. 12:19PM

24 Q. (By Mr. Mills) Do you recall being asked 12:19PM

25 questions about size selection involving the removal of 12:19PM

1 dimers? 12:19PM

2 A. Yes. 12:19PM

3 Q. Is size selection limited to the removal of 12:19PM

4 dimers? 12:19PM

5 MR. POWERS: Objection. Leading. 12:19PM

6 A. No, certainly not. 12:19PM

7 Q. (BY Mr. Mills) Please take a look at 12:19PM

8 paragraph 23 of your declaration. 12:19PM

9 MR. POWERS: Can you give me a minute to 12:19PM

10 get there? 12:19PM

11 MR. MILLS: Yes. 12:19PM

12 Q. (By Mr. Mills) In paragraph 23 of your 12:19PM

13 declaration you cite page 35 of Exhibit 1035. Correct? 12:19PM

14 A. Yes. 12:20PM

15 Q. Can you tell us what Exhibit 1035 is? 12:20PM

16 A. Exhibit 1035 is the template prep sequencing 12:20PM

17 guide. 12:20PM

18 Q. In paragraph 23 of your declaration you quote 12:20PM

19 page 35 of Exhibit 1035 as saying: It is highly 12:20PM

20 recommended to select amplicons of the same size, plus 12:20PM

21 or minus 10 percent to minimize loading bias. 12:20PM

22 MR. POWERS: Objection. Leading. 12:20PM

23 Is that even a question? 12:20PM

24 A. Yes, that is what I said. 12:20PM

25 Q. Dr. Larrea, do you recall being asked 12:20PM

1 questions earlier about polymerase reads and subreads? 12:20PM

2 A. Yes. 12:21PM

3 Q. Is a subread a sequence read? 12:21PM

4 MR. POWERS: Objection. Leading. 12:21PM

5 A. Yes. 12:21PM

6 MR. MILLS: No further questions. 12:21PM

7 MR. POWERS: Okay, let's go off the record 12:21PM

8 two minutes. 12:21PM

9 (Recess taken 12:21 p.m. to 12:22 p.m.) 12:21PM

10 * * * 12:21PM

11 12:22PM

12 RE CROSS-EXAMINATION 12:22PM

13 BY MR. POWERS: 12:22PM

14 Q. Dr. Larrea, at the break that we took before 12:22PM

15 your redirect, did you speak with your counsel about 12:22PM

16 the substance of your testimony that you gave on 12:22PM

17 redirect? 12:22PM

18 A. I'm not sure what you mean by "substance," but 12:22PM

19 we did discuss some of the general concepts. 12:23PM

20 Q. Did counsel indicate to you the questions that 12:23PM

21 counsel would be asking? 12:23PM

22 A. The exact questions were not given. The 12:23PM

23 general topics were discussed. 12:23PM

24 Q. Did counsel discuss the topics of your 12:23PM

25 redirect testimony with you during the break? 12:23PM

1 Q. (By Mr. Powers) The type of answer that you'd 12:26PM
2 give to those questions was discussed? 12:26PM

3 MR. MILLS: Objection to form. 12:26PM

4 A. In general terms, yes. 12:26PM

5 Q. (By Mr. Powers) The substance of the answer 12:26PM
6 that you would give was discussed. Correct? 12:26PM

7 A. What do you mean by the "substance"? 12:26PM

8 Q. The type of answer that you would give. 12:26PM

9 A. Right. And so I said the type of answer that 12:26PM
10 I would give, yes. 12:26PM

11 Q. That was discussed with counsel during the 12:26PM
12 break? 12:26PM

13 A. That was discussed. 12:26PM

14 Q. Thank you. 12:27PM

15 MR. POWERS: We can go off the record. 12:27PM

16 (Deposition concluded at 12:27 p.m.)

17 (Signature was reserved.)

18 * * *

19

20

21

22

23

24

25

REPORTER'S CERTIFICATE

I VICKY L. PINSON, RPR-CCR, the undersigned
Certified Court Reporter, authorized to administer
oaths and affirmations in and for the states of
Washington, Oregon and California, do hereby certify:

That the sworn testimony and/or proceedings, a
transcript of which is attached, was given before me at
the time and place stated therein; that the witness was
duly sworn or affirmed to testify to the truth; that
the testimony and/or proceedings were stenographically
recorded by me and transcribed under my supervision.

That the foregoing transcript contains a full,
true, and accurate record of all the testimony and/or
proceedings occurring at the time and place stated in
the transcript; that a review of which was requested.

That I am in no way related to any party to the
matter, nor to any counsel, nor do I have any financial
interest in the event of the cause.

WITNESS MY HAND this 9th day of May, 2023.



VICKY L. PINSON, RPR-CCR

Washington Certified Court Reporter, No. 2559

Oregon State Certified Court Reporter, No. 16-0442

California State Certified Court Reporter, No. 9845.

1 JAD A. MILLS

2 jmills@wsgr.com

3 May 11, 2023

4 GUARDANT HEALTH, INC. v. University Of Washington Et Al.

5 5/9/2023, WITNESS (#5891566)

6 The above-referenced transcript is available for
7 review.

8 Within the applicable timeframe, the witness should
9 read the testimony to verify its accuracy. If there are
10 any changes, the witness should note those with the
11 reason, on the attached Errata Sheet.

12 The witness should sign the Acknowledgment of
13 Deponent and Errata and return to the deposing attorney.
14 Copies should be sent to all counsel, and to Veritext at
15 clientservices-yom@veritext.com

16
17 Return completed errata within 30 days from
18 receipt of testimony.

19 If the witness fails to do so within the time
20 allotted, the transcript may be used as if signed.

21
22 Yours,

23 Veritext Legal Solutions
24
25

1 GUARDANT HEALTH, INC. v. University Of Washington Et Al.

2 ANDRES LARREA, Ph.D. (#5891566)

3 E R R A T A S H E E T

4 PAGE_____ LINE_____ CHANGE_____

5 _____
6 REASON_____

7 PAGE_____ LINE_____ CHANGE_____

8 _____
9 REASON_____

10 PAGE_____ LINE_____ CHANGE_____

11 _____
12 REASON_____

13 PAGE_____ LINE_____ CHANGE_____

14 _____
15 REASON_____

16 PAGE_____ LINE_____ CHANGE_____

17 _____
18 REASON_____

19 PAGE_____ LINE_____ CHANGE_____

20 _____
21 REASON_____

22 _____
23 _____
24 ANDRES LARREA, Ph.D.

Date

1 GUARDANT HEALTH, INC. v. University Of Washington Et Al.
2 ANDRES LARREA, Ph.D.(#5891566)

3 ACKNOWLEDGEMENT OF DEPONENT

4 I, ANDRES LARREA, Ph.D., do hereby declare that I
5 have read the foregoing transcript, I have made any
6 corrections, additions, or changes I deemed necessary as
7 noted above to be appended hereto, and that the same is
8 a true, correct and complete transcript of the testimony
9 given by me.

10
11 _____
12 ANDRES LARREA, Ph.D. Date

13 *If notary is required

14 SUBSCRIBED AND SWORN TO BEFORE ME THIS
15 _____ DAY OF _____, 20____.

16
17
18 _____
19 NOTARY PUBLIC
20
21
22
23
24
25

[& - 4]

&	60:13,15,16,19	2	202.371.2600
& 2:5,13 4:14	1036 8:10 45:16	2 3:10 8:5	2:14
0	10:07 34:11	28:23,24 34:20	2021 8:15
0.4kb 45:5,13	10:24 34:11	39:23 40:13	2023 1:20 4:1
0.4kb. 45:10	11 8:15 65:3	41:3 50:5 52:1	64:20 65:3
075 27:15	1100 2:13	55:13,17	2026 8:17
33:19 48:15,19	11:10 53:21	2,000 46:24	206.883.2500
53:17,22 56:6	11:19 53:21	47:14	2:7
56:10,15,20,23	11:32 59:9	20 67:15	21359 64:23
57:2 58:12	12 8:17 46:10	20005 2:14	23 54:7 60:8,12
09 50:1	46:13 47:18	200090024331	60:18
1	12:18 59:9	56:24	24 43:20,20
1 3:9 8:3 10:17	12:21 61:9	2001 8:11	44:8,15,18,25
14:6,7 27:22	12:22 61:9	2002 8:12	2559 1:24
28:17 54:1,13	12:27 63:16	2004 50:1	64:24
54:19	13 47:12	2006 8:13	28 3:10
1,350 50:21	138 54:2	2009 8:11	2d 52:21,21
51:5	139 54:2,4	2009/002433...	3
10 8:14 27:3,10	14 3:9	3:25	3 3:11 8:6 30:2
60:21	141 56:20	20090024331	30:3 31:19
10,760,127 1:15	145 48:21	57:4,9 58:11	32:10,20,23
100 47:21 48:4	55:24	2009029807...	33:2,6,9 43:4,5
1000 48:23	15 27:8,10	8:5	45:24,25 55:20
49:3,5,10	52:13	2010 8:12,16	55:20 56:6
50:19,20,23	150 56:9,15	27:14,17 28:4	30 3:11 52:11
51:3 56:16	1500 55:21	28:10,12,14,17	52:12,13 65:17
1018 8:5 48:19	16-0442 1:24	33:1,5,19	300 50:24 51:4
53:23	64:24	34:19,20,23	31 3:14,17
1020 8:6	162 55:14	35:1,5,19,20	34 43:19
1021 8:7 27:18	17,000 47:13	46:21	35 60:13,19
1033 8:3	181 57:21 58:7	2011 8:10	39 3:20
1034 23:16	58:10,17	2012 8:13	4
1035 8:8 43:1	1a 28:17	2014 8:14	4 3:3,14 8:7 31:5,6,19
43:11 44:8,15		2015 4:22	

[4 - appended]

32:10 35:1 56:9,10 49 3:21	98104-7036 2:6 9845 1:24 64:25 9:06 1:20 4:2 9th 64:20	adapters 38:17 56:13 adaptors 38:10 additions 67:6 address 11:18 addressed 16:2 adequately 45:3 administer 64:4 adoption 19:20 advantage 21:20,22 affirmations 64:5 affirmed 64:10 agarose 52:11 52:21 ago 13:11 agree 20:9,17 28:13 29:3,9 29:21 33:18,21 35:5 37:2,8,15 38:1,9,16 39:3 39:14 40:9 41:22 42:19 51:11,15,19 54:23 55:13 56:6 agreed 18:20 agreement 26:15,16 ahead 6:5 14:5 48:12	al 65:4 66:1 67:1 alleged 9:15,20 9:25 57:12 allotted 65:20 allowed 19:10 amount 45:11 amplicons 60:20 ampure 43:21 43:25 44:4,6,9 44:15,19,20,21 44:22 45:9,11 47:24 48:1,3 analysis 13:2 57:11 andres 1:17 4:5 66:2,24 67:2,4 67:12 annealing 31:21 answer 6:5,19 11:5,25 12:16 12:19 21:14 62:6,10,15 63:1,5,8,9 answered 11:9 13:8 21:11 23:1,10 answers 6:1,15 26:13 62:7,18 62:22,25 appeal 1:3 appended 67:7
5	a		
5 3:17 8:8 31:24,25 32:10 5/9/2023 65:5 50 52:11 500 55:21 51 3:23 5100 1:21 2:6 57 3:24 5891566 65:5 66:2 67:2 59 3:4	a.m. 1:20 4:2 34:11,11 53:21 53:21 59:9 abide 26:15 above 65:6 67:7 acceptable 6:16 accident 5:7,16 accomplishes 21:22 accomplishing 18:14 accuracy 65:9 accurate 6:15 13:17,19 23:17 64:14 achieve 22:9 achieved 20:22 21:5,6 23:5 achilles 5:18,20 acknowledge... 67:3 acknowledg... 65:12 actually 53:18 actuators 3:23 adapter 42:11 45:5		
6			
6 3:20 8:10 17:7 39:21,22 40:13 50:18 60 52:12 61 3:5			
7			
7 3:21 8:11 49:13,14 7,476,503 8:6 701 1:21 2:6			
8			
8 3:23 8:12 51:24,25			
9			
9 1:20 3:24 4:1 8:13 57:18,19 58:3			

[applicable - break]

applicable 65:8	22:10	backgrounder	32:11,15 35:21
application	artisans 18:13	8:16	36:3,15
3:24 26:22	18:21 20:10	bands 52:13,24	bio 24:20,21
56:24 57:3,8	aside 9:4 14:17	base 41:5 46:24	25:2,6,9,11
57:14 58:10	asked 10:24	47:21 48:4	biobank 24:22
applications	11:8 13:7	50:21,24 51:4	24:24
24:7	21:11 23:1,10	51:5 55:21	bioscience's
applied 27:21	59:15,24 60:25	based 13:2 19:9	42:16
48:23	asking 11:11	21:14 32:18	biosciences
appreciate	21:4 23:5	47:23	3:20 25:8,10
17:25	25:25 61:21	bases 39:4	26:7,10,14,19
approach 18:1	assignee 57:23	basis 12:14,18	26:21 27:2,10
20:18,23,25	58:4	33:17 40:3	40:7 44:2,5
21:7,9 22:11	assume 6:6	49:20 52:6	57:24 58:4
23:8	assumption	bead 43:7	biotin 35:8,11
approaches	41:1	47:23	35:14 36:22
20:15,23 21:6	attach 36:4	beads 43:11,22	37:1,3,5
22:21	attached 36:8	43:25 44:4,6,9	biotinylated
appropriated	36:10 43:17,18	44:16,20,20,21	34:23
53:8	64:8 65:11	44:22 45:9,12	bit 5:6 25:3
approximate	attorney 65:13	48:1,4	45:1
4:21	authenticity	beginning 7:24	blowup 28:22
approximately	49:20	behalf 4:14	board 1:3
4:22 46:14	authorized	believe 5:3,15	bottom 35:13
47:5	64:4	23:18 35:16	36:11,23 43:5
area 29:17	available 65:6	best 44:14	45:25 47:19
arriving 14:18	avenue 1:21 2:6	bias 60:21	50:4
15:3	2:13	binder 7:24	bound 32:15
arrow 32:7	b	10:2 27:13	34:6 35:12
arrows 31:17	b 3:23	43:2 45:16	36:14,16 38:7
articulated	back 10:1	48:16	41:7,14 46:2
36:13	45:19 46:12	binding 29:14	box 5:17
artisan 17:13	47:18	29:18,22 30:10	bp 47:14
17:17,25 20:17	background	30:17,18 31:12	break 34:9,14
20:24 21:8,25	41:7	31:18,21 32:5	34:17 48:10

[break - concluded]

53:19 59:4 61:14,25 62:3 62:4 63:12 brief 27:22 broad 41:10 bromide 53:9 53:10 brought 7:3 10:11 bucket 41:10 building 24:22	catching 57:25 cause 64:19 caution 26:15 cautionary 12:15 26:12 ccr 1:24 64:3 64:23 cell 35:19,20,21 52:22 celsius 54:7 center 1:9 certain 11:17 certainly 60:6 certificate 64:1 certified 4:6 64:4,24,24,25 certify 64:6 cetera 45:4 challenge 22:14 change 66:4,7 66:10,13,16,19 changes 65:10 67:6 cheaper 22:21 check 52:23 chemical 3:23 chin 8:10 13:12 48:3 choose 20:18 20:25 21:9,25 22:10,20 23:8 chroma 3:21 48:23 49:3,5,9 50:13,19,20,23 51:3 56:16	circle 29:14,17 30:9,16 circular 27:21 29:4,6 35:3 37:18 38:3,11 38:18 cite 14:22 15:3 60:13 cited 7:5,9,12 9:1 15:15,17 15:20,21,22,24 58:11 cites 46:21 47:24 clarification 36:17 44:19 47:9 51:6 class 59:22 cleanup 54:14 clear 15:1 33:5 clientservices 65:15 clone 50:5,9 51:2 close 54:24 cmw 33:18 coating 35:6 coded 31:2 coffee 48:11 coli 55:21 collaborative 14:2,3 collected 41:4 college 23:24 24:1	color 31:2,14 colors 30:22 column 35:2 48:24 56:17 columns 3:21 49:3,6,10 50:13,19 come 38:21 comments 13:21,21,25 commercializ... 1:9 commonly 20:20 company 24:22 comparable 20:17 22:21 complete 6:15 67:8 completed 65:17 complex 36:10 36:19,23 complexes 41:6 41:14 comprised 27:10 43:14 comprising 46:1 concentrated 47:20 48:22 concepts 61:19 concluded 63:16
c			
c 2:1 calculated 22:5 california 1:24 57:24 58:5 64:6,25 call 10:12 calling 12:21 41:5 calls 12:17 capped 46:19 carcinogen 53:11 carried 54:12 cartoon 32:18 case 1:14 4:24 5:1,4 8:21 9:7 9:12,14,17,23 10:8,21 11:13 13:15 14:9,13 23:16 57:10 cases 22:15 cassettes 19:10 19:14			

[conditions - declaration]

conditions 18:7 43:8 48:7 55:5 conductive 41:5 confidential 12:17 confirm 58:9 59:6 consequently 52:20 consider 9:5 14:22 15:2 25:20,25 constructed 45:25 construction 54:3 constructs 46:1 consultant 24:15 consulting 24:16 consuming 51:16 52:25 contain 42:8 contained 48:24 56:1 containing 55:14 contains 64:13 contents 3:9 7:23 8:1 10:1,2 10:16 11:7,14 11:20 12:2,25 13:6 14:5	context 17:19 40:6 continue 34:19 35:4 continues 46:23 52:19 53:6 control 32:14 controlled 26:14 convenient 42:24 conversations 12:18 copies 7:4 65:14 copy 7:4 45:20 correct 5:2,3 7:6,11 8:21,22 10:19 14:23 16:2 17:5,6 18:9,15,22 20:1,4 24:11 24:13,17 26:23 26:24 28:4,10 29:7 31:19 32:12,16,18,24 32:25 33:25 34:8,24,25 35:7,14,16 36:5,6,9 40:19 41:8 42:5 43:12,13 44:10 44:16 45:10,14 45:15 46:8,21	47:2,3,15,24,25 48:6 49:7 50:1 50:6,21,25 51:5 54:10,13 54:17,20 55:8 55:10,18,22,23 56:4,10,13,14 56:18 60:13 63:6 67:8 corrections 67:6 correctly 42:12 42:15 43:9 45:7 46:5 52:15 53:2,12 53:15 59:2,3 cost 20:5 22:15 23:2 costly 22:1 23:8 costs 17:10 19:4,6 20:4,11 21:7,24 22:4,7 22:8,11 counsel 6:18 10:25 11:6,13 11:19 12:1,4 12:18,23 13:5 13:15 14:4,17 16:9,12 17:4 34:13,16 61:15 61:20,21,24 62:2,5,6,9 63:11 64:18 65:14	court 4:6 6:10 35:9 36:17 64:4,24,24,25 covaris 46:25 created 54:10 cross 3:3 4:9 currently 24:17 25:7 59:21 cut 17:9 19:4,6 20:3,11 cutting 20:5 28:3,10,12,13 cv 23:13,17 25:12
			d
			d 3:1 data 33:10 38:8 41:17 45:18 46:12 58:23 59:1 date 66:24 67:12 david 2:12 4:15 day 16:24,25 17:1 64:20 67:15 days 65:17 dc 2:14 deal 18:18 20:13 21:2 22:17 23:11 declaration 6:25 7:2,4,5,10 7:13 8:4,14,20 8:23,24 9:2,6

[declaration - drivers]

9:10,17,19,23 10:8,9,13,18,21 10:24 11:16 13:1,15,20,22 13:23,25 14:1 14:11,16,19,23 15:3,10,11,16 15:18,22,24 16:1,2 17:7 41:25 57:13 60:8,13,18 declare 67:4 deemed 67:6 defendant 5:10 5:11,14 definition 25:23,24 40:13 42:17,19,22 degrees 54:7 delivering 5:18 depend 17:22 17:23 18:25 20:21 22:4,17 23:11 39:18 54:22 dependent 51:20 52:25 depending 55:4 depends 17:21 18:7,18 20:13 21:2 45:11 deponent 65:13 67:3 deposed 4:17 4:18 5:22	deposing 65:13 deposition 1:17 7:4 10:12 15:6 15:18 16:6 59:16 63:16 depot 5:15 derivative 44:20 describe 25:15 38:23 described 27:22 28:6 33:18 35:19 36:3 46:3 describing 42:24 description 13:18,19 design 18:19 19:1 22:18 23:4 designed 19:15 25:17 details 21:14 21:16 developed 27:23 28:8 developing 26:20 27:1 dholman 2:15 different 19:15 20:23 21:6 differently 10:6 difficult 59:18	digested 54:5 dimer 43:14 dimers 43:6,12 45:6 60:1,4 direct 27:18 28:17 34:20 40:12 57:20 directly 36:7 disagree 42:16 42:21 disclose 33:1,6 disclosed 33:11 discloses 43:11 44:8,15 48:3 49:5 disclosure 12:17 discuss 34:12 34:16 61:19,24 62:4 discussed 61:23 62:2,17,21,23 63:2,6,11,13 discusses 54:3 discussion 62:11,14 dissolve 52:21 distribution 33:13,20 47:8 distributions 47:13 dna 36:4 37:17 38:2 46:1,2,18 46:23 50:20,24 51:3,4 52:13	52:24 53:7 54:6 doctor 52:9 document 14:14 27:19 39:25 40:2 43:4,19 49:19 49:22,25 50:12 50:18,19 52:6 52:19 53:6 57:2 documents 6:22,23,24,25 7:15,18,21 8:19 9:4 10:3,7 10:10,16,17,20 11:1,6,13,20 12:1,24 13:1,5 13:16 14:13 15:12,17,19,21 15:25 27:14 36:14 double 14:25 46:18 dr 4:11 9:14,16 9:20,22 13:20 13:25 14:11,15 15:13,20,25 23:17 30:1 34:12 49:23 57:12 59:14 60:25 61:14 drawing 29:5 drivers 22:16
--	--	---	---

[due - familiar]

due 41:6 56:1	ends 46:20	excerpts 9:18	experience
duly 64:10	entire 8:23,24	excess 54:6	25:12
dyes 53:9	19:11,13	excuse 58:20	experiment
e	entirety 57:5	exhibit 3:9,10	17:15,23 19:23
e 2:1,1 3:1	enzyme 58:23	3:11,14,17,20	20:1,2,14 21:3
55:21 66:3,3,3	58:25 59:1	3:21,23,24	39:19 52:9
earlier 61:1	enzymes 45:4	7:25 8:3,5,6,7,8	53:10 54:18
edge 28:3,10,12	equal 17:24	8:10,11,12,13	55:11
28:13	18:6 23:7	8:14,15,17	experimental
efficiencies	equivalent 22:9	10:17 14:6,7	18:7,19 20:21
17:9	errata 65:11,13	23:16 27:18	22:18 23:4
efficiency	65:17	28:23,24 29:13	experiments
17:13,16 18:8	et 45:4 65:4	30:2,3 31:5,6	18:24 25:18
18:11,20,24	66:1 67:1	31:24,25 39:21	expert 4:23
19:4,6,12 20:1	ethanol 48:22	39:22 40:13	24:3,5 25:19
20:10 22:15	ethidium 53:9	43:1,11 44:8	25:22,23,25
23:2	53:10	44:15 45:16	42:4
efficient 18:1	event 5:21	48:19 49:13,14	expressed
18:14 20:16,18	64:19	51:24,25 53:23	14:18
20:24,25 21:7	events 38:22	57:18,19 58:3	extent 12:16
21:23 22:1,8,9	exact 4:20 14:1	60:13,15,16,19	26:13
22:11,22,23	18:18 20:13,21	exhibits 3:8 7:5	f
23:8 41:5	22:17 27:7	7:7,8,9,12	fact 5:1 59:21
effort 14:2,3	39:19 61:22	10:23 31:19	factors 21:13
egan 8:13	exactly 11:10	32:10	23:3,3,12
eid 8:11,17	54:20,21	exists 26:16	faculty 23:24
13:12	examination	exonuclease	24:1
eight 17:2	3:3,4,5 4:9	56:2,3	failed 55:25
either 23:6	59:12 61:12	exonucleases	fails 65:19
electrophoresis	example 13:9	45:4	fair 13:4,16
51:11,16 52:4	19:7,9 21:19	expected 58:20	familiar 11:3
52:12,22,22,23	54:1,13,19	58:21	28:25 37:10,13
59:16,19	55:13,17,20,20	expensive	38:24 39:10,13
enable 27:24	56:6,9,10	22:22	40:5 50:9

[familiarity - goes]

familiarity 25:15	12:18 13:7 14:24 16:3 17:14 18:2,5 18:10,16,23 20:12,19 21:1 21:10 22:2,12 22:25 23:9 28:5 29:19,23 32:17 33:4 35:15,23 36:18 37:12 41:24 42:8,13,18,23 51:7,13,17 52:16 53:4,13 62:24 63:3	full 6:15 64:13 function 33:16 fundamentally 36:25 funded 20:7 further 41:19 61:6 furthermore 53:7	18:21 19:3 genetics 59:22 genome 19:11 19:14,16 55:21 genomic 46:23 gerrard 2:5 give 6:15 23:15 49:16 60:9 62:18,22 63:2 63:6,8,10
fifth 1:21 2:6 figure 28:17,17 28:23 30:21 32:20,23 33:2 33:6,9 filed 8:20 9:11 9:13,17,23 10:8,13 24:8 financial 64:18 fine 12:10 finish 12:7 first 5:25 9:16 16:13,14 28:16 35:2 45:1 49:25 57:22,23 58:3 five 23:21 31:11 52:12,14 floor 36:16 florida 24:23 flow 35:19,20 35:21 focus 9:14 focused 8:25 9:19,24 14:10 14:15 57:11 fold 54:5 follow 33:19 follows 4:7 foregoing 64:13 67:5 form 9:8 10:4 10:22 11:8	format 46:8 formed 43:6 foundation 21:18 49:20 50:2 51:8 58:13 four 3:15 9:15 9:20,25 14:15 15:13,20,21,23 16:1 27:19 31:9 57:11 fox 2:13 4:14 fragment 47:13 fragments 47:4 47:21 48:4 55:22,24 front 6:23 7:16 9:4 23:13 27:13 29:13 48:18 57:22	g galatea 24:19 24:21 25:2,6,9 25:11 gel 51:11,15 52:4,11,13,21 52:23,24 59:16 59:18 gen 17:19 18:11 19:10,21 19:24 general 61:19 61:23 62:17 63:4 generally 17:16 20:10 generate 40:21 55:21 generated 3:12 3:15,18 30:5 30:14 31:9 32:2 40:16 41:17 47:5 54:16 55:14 generation 17:8 18:3,8,15	given 31:21 42:19 61:22 62:7 64:8 67:9 giving 21:14 glossary 3:20 42:6 go 5:23 6:5 10:1 14:5 17:7 39:23 43:1,4 43:19 46:10 47:12,18 48:12 48:12,21 49:15 50:18 52:1 55:13,20 56:9 58:9,17 61:7 63:15 goal 17:23 18:8 18:25 19:25 20:3,21,22 21:5,5 23:5 33:21,22 goals 18:15 21:23 22:9 goes 30:21 35:1

[going - inserts]

going 3:12,15 3:18 12:8,14 26:11 30:1,6 30:15 31:9 32:3 34:10 38:22 39:20 45:23 46:10 51:23 57:17 58:7 59:5 goldstein 2:13 4:14 good 4:11,12 goodrich 2:5 great 18:18 20:13 21:2 22:17 23:11 48:18 greater 47:21 48:4 50:21 51:4 green 31:1 ground 5:24 guardant 1:5 65:4 66:1 67:1 guess 11:10 35:25 46:10 guide 8:9 13:10 60:17	hairpins 43:15 43:17 halfway 45:2 hand 13:10 28:16,22 30:1 31:4,23 39:20 51:23 57:17 64:20 handled 62:12 62:14 happens 44:14 happy 37:25 hardware 26:6 heading 34:21 health 1:5 65:4 66:1 67:1 hearsay 40:3 49:20 52:7,17 53:5,14 hereto 67:7 high 41:7 59:22 highlighted 12:4 highly 60:19 history 14:12 hmm 10:14 26:5 41:21 holds 36:23 holman 2:12 4:15 home 5:15 hope 5:19 hour 16:13,20 34:10 54:7	hours 17:2,5 house 5:7 25:4 25:5 hughes 20:7 huh 27:16 hypothetical 21:5,12 22:7 62:10	incomplete 56:1 incorporated 37:17 38:2,7 57:4 58:14 incorporates 56:23 incorporation 38:22 increase 19:19 increased 19:17 incubating 54:5 indicate 31:18 61:20 62:5,9 indicating 7:19 29:20 31:3,16 indirectly 36:14 individual 38:21 inefficient 21:19 information 12:22 initial 54:11 initially 24:12 innovate 26:6,9 26:18 insert 42:10 43:16 49:6 inserted 19:15 inserts 48:25 48:25 49:7
h	h	i	
h 2:12 66:3 hairpin 31:1 43:6,12,14 46:19 54:6 55:25 56:12		i.e. 47:21 idea 31:7 identification 3:8 14:7 28:24 30:3 31:6,25 39:22 49:14 51:25 57:19 identified 11:17,18 iii 2:12 4:13 56:2 illumina 25:7 25:13,16,18,19 25:21 26:1,3 imaging 52:13 52:23 immobilization 47:23 important 6:9 17:13,17 18:21 19:25 20:3,5 improve 17:9 19:4,5,12 20:1 20:10 include 6:24 38:10,17 40:21	

[inspect - low]

inspect 7:18 instruct 12:19 instructed 6:19 instruction 12:15 26:12 instructions 47:1 instruments 52:20 intensive 51:12 52:25 interaction 35:11,14,17 36:9,22 37:1,2 37:5,9 interactions 37:4 interest 64:19 interesting 44:23 intervening 43:16 introduce 19:3 introduced 17:8 inventor 24:8 involves 52:10 involving 59:25 ipr2022-00817 1:14 issue 12:15 14:9 26:11 issued 26:23	j jad 2:4 16:11 65:1 jmills 2:7 65:2 job 6:11 juniors 59:23	51:24 57:18 59:14 60:25 61:14 66:2,24 67:2,4,12 lawsuit 5:5,8,9 5:14 leading 59:20 60:5,22 61:4 62:23 legal 65:23 legend 30:21 31:20 length 38:25 39:3,7 54:20 54:21,24 lengths 42:2 libraries 46:1 library 26:20 27:1 55:3 ligase 54:7 ligases 45:4 ligate 55:25 ligated 55:7 56:12 ligating 55:9 ligation 45:5,9 48:22 54:4,22 55:5 56:1 light 53:8 limited 60:3 line 32:6 66:4,7 66:10,13,16,19 linear 3:11,14 3:17 29:7,8,9 29:12 30:5,14	31:8 32:2 lines 47:19 linkers 31:1 listed 7:25 9:5 11:7,13,20 12:2,24 13:6 little 5:6 10:6 45:1 loaded 33:24 36:20 loading 33:17 33:18,21,22 34:2,5 52:11 60:21 located 24:22 58:24,25 locations 58:19 58:22 long 16:12,18 16:23 37:20 look 19:11 32:10 41:19 44:18 49:25 50:4 56:20 57:22 58:3 60:7 looked 58:12 looking 18:14 19:13 30:12,13 loops 46:19 lot 52:20 louder 35:10 47:10 low 39:5,8
	k kessler 2:13 4:13,16 know 5:23 6:4 26:2 33:12 43:25 44:4 48:1 49:3,9 knowledgeable 25:20		
	l l 1:24 64:3,23 lab 19:4 labor 51:11 52:25 laboratories 50:5,10 lacking 40:19 40:24 language 28:1 46:23 52:9 large 24:25 58:18 larrea 1:17 4:5 4:11 8:3 10:17 14:6 23:17 28:23 30:1,2 31:5,24 34:12 39:21 49:13,23		

[made - negligence]

m	mathematical	21:10,18 22:2	molecules
made 13:25	33:16	22:12,25 23:9	27:21 35:3
44:4,6 67:5	matter 64:18	26:11 28:5,11	50:20,24 51:3
magnetic 43:7	maximize 34:1	29:11,19,23	51:4 53:7
43:11	mean 10:9 34:6	32:17 33:4,8	moment 39:25
make 7:20	47:4 61:18	35:15,23 37:12	44:13 49:17
37:23,23 44:21	62:14 63:7	37:19 38:5,12	52:2,5
55:1 56:13	means 40:9,17	38:19 39:6,12	monitored
makes 56:10	41:3,3 47:12	39:17,24 40:2	58:20
manner 34:16	meant 41:17	40:11,14,20,25	morning 4:11
46:3	mechanically	41:9,15,24	4:12
manual 3:22	46:24 47:7	42:13,18,23	motivated
manufacturer	meet 16:5,8,12	47:6,16 49:16	20:10,25 21:8
49:9 50:13	16:16,18,21,23	49:19 50:2,7	22:10 23:7
51:2	memorialized	51:7,13,17,21	moving 54:2
manufacturer's	13:22	52:2,5,16 53:4	multiple 38:11
47:1	met 16:9,11,13	53:13,20 58:13	38:18 41:6,13
mark 10:16	17:4	59:13,24 60:7	multiplex
14:5 28:23	method 22:1,7	60:11,12 61:6	27:24
30:2 31:5,12	34:1	62:24 63:3	n
31:23 32:4,6	methods 26:10	65:1	n 2:1 3:1
39:21 49:12	26:18	minimize 60:21	name 16:10
51:24 57:17	metric 22:6	minus 60:21	narrow 55:4
marked 14:7	miami 24:23	minute 60:9	nascent 28:3,10
23:16 28:24	microwave	minutes 52:11	nature 5:4,16
30:3 31:6,17	52:21	52:12,12,13,14	37:4
31:25 39:22	mike 16:11	61:8	neb 54:7
49:14 51:25	mills 2:4 3:4	missing 42:2	necessary 67:6
57:19	9:8 10:4,22	mode 33:23	need 6:1,14,18
markings 7:20	11:8,15,22	34:6,8 35:6,13	21:17 39:25
material 8:18	12:3,5,8,11,13	36:4,8,15,16	45:22 48:11
9:3,9 11:18	13:7 14:24	39:4 40:10,17	negative 14:25
13:2,13 55:25	16:3 17:14,18	41:4	negligence 5:8
materials 9:1,5	18:2,5,10,16,23	molecule 27:24	5:9
	20:12,19 21:1	29:7,8,10,12	

[neither - page]

neither 41:11	object 6:18	occurring	owner 1:10
never 24:1,5	18:16 40:2	64:15	2:11
new 2:13 13:13	49:19 52:6	offered 13:20	p
26:6,9,18,20	objection 9:8	office 1:1	p 2:1,1 41:3
27:1	10:4,22 11:8	okay 5:6,19,22	p.m. 59:9 61:9
nick 56:1	12:9,13 13:7	6:4,7,8,14,22	61:9 63:16
niedringhaus	14:24 16:3	10:25 11:25	p0 40:6,17,21
8:12	17:14,18 18:2	13:4,14 16:16	41:11
nih 20:8	18:5,10,23	23:13,15 25:15	p1 40:6,9,15
noncovalent	20:12,19 21:1	27:6,9 31:17	41:11
37:4	21:10,18 22:2	32:9 36:2,12	p2 40:6 41:10
nondisclosure	22:12,25 23:9	40:4 42:21	41:13,16
26:15	28:5 29:11,19	44:25 48:18	pacbio 8:15
notary 67:13	29:23 32:17	53:25 57:17	11:2 23:22,23
67:19	33:4,8 35:15	61:7	24:10,12 25:12
note 65:10	35:23 37:12,19	older 19:21	28:12 33:17,23
noted 49:21	38:5,12,19	oligo 54:6	36:7,13,18
67:7	39:6,12,17	once 4:18 31:14	37:11 38:8,20
notes 7:20	40:11,14,20,25	38:14 39:8	38:25 39:11,15
nucleotide	41:9,15,24	ones 43:2	39:18 41:22
36:20	42:13,18,23	opinion 14:18	42:4 44:7 46:4
nucleotides	47:6,16 50:2,7	15:4	pacbio's 36:2
37:16 38:2,6	51:7,13,17,21	opinions 9:6	42:6
38:21 39:8	52:16 53:4,13	12:19 57:9,15	pacific 3:20
55:15	58:13 59:20	opposed 19:21	25:8,10 26:7
number 21:12	60:5,22 61:4	ordinary 22:20	26:10,14,18,21
23:11 27:7,23	62:24 63:3	23:7	27:1,9 40:6
39:4	objections	oregon 1:24	42:16 44:2,5
numbers 7:25	11:15,22 12:3	64:6,24	57:23 58:4
58:19	28:11	outcome 55:10	page 3:2,8
nw 2:13	observation	outside 18:16	27:19,20 34:20
o	27:25	22:2,12,25	35:1 39:23
oath 6:14	obtain 33:23	23:9 42:13	40:13 41:19
oaths 64:5	occur 58:23	52:7	43:4,5,19,20
			44:8,15,18,25

[page - powers]

45:24,25 46:10 46:13,13 47:12 47:18 49:25 50:5,18 52:1 57:22,23 58:3 60:13,19 66:4 66:7,10,13,16 66:19 paid 24:14 pair 55:22 pairing 52:10 pairs 46:25 47:22 48:4 50:21,25 51:4 51:5 paper 33:10 papers 15:15 paragraph 17:7 27:20 35:2,5 45:1,2 46:13,18 48:21 52:3,8 54:2 55:24 56:9,15 56:20 57:21 58:7,10,17,18 60:8,12,18 paramagnetic 47:23 parameter 18:11 part 11:19 13:1 13:14 14:13 25:1 26:20 particular 17:15 19:7,9	19:23 20:6 58:24 particularly 51:14,18,22 partitioned 39:15 41:23 42:8 parts 19:15 37:23 38:23 party 64:17 pass 42:9 59:5 passage 28:7 passed 56:16 passes 38:11,18 past 50:11 patent 1:1,3,10 1:15 2:11 3:24 14:8,10 24:7 26:22,23 56:24 56:24 57:3,8 patents 24:7,9 pause 12:4 29:24 35:24 pb 43:21,25 44:2,4,6,9,16 44:19,21 45:9 pcr 54:5,17,25 55:1,4,6,9 56:10,12 peer 23:19 pen 29:15 people 13:10 26:4,25 27:6 27:10	percent 60:21 percentage 58:22 perform 45:2 59:18,23 performed 54:4 person 22:20 23:7 petition 9:11 9:13 petitioner 1:6 2:3 ph.d. 1:17 2:12 2:12 4:5 66:2 66:24 67:2,4 67:12 phase 47:22 phrase 38:14 pi 20:6 pinson 1:24 64:3,23 place 64:9,15 plaintiff 5:12 platform 25:7,9 25:11 platforms 25:6 play 21:13 23:4 please 7:24 12:6 30:9,20 32:4 58:2 60:7 pllc 2:13 plus 60:20 point 44:19 poisson 33:12 33:19,22	polymerase 3:12,15,18 29:14,18,21 30:6,10,15,17 30:18 31:9,12 31:18 32:3,5 32:11,15 34:6 36:9,15,19 37:10,13,16,17 38:1,3,6,7,9,13 38:16,21,25 39:3,7,15 40:10,15,19,24 41:2,6,14,23 42:2,7,9 46:2 61:1 polymerases 34:23 35:12 poor 53:1 portions 31:1 42:25 posed 62:6 possible 18:14 33:24 possibly 41:6 potential 30:18 31:21 potentially 7:12,14 power 52:22 powers 2:12 3:3,5 4:10,13 12:6,10,12,21 14:8,25 16:5 17:16,19 18:3
---	--	---	---

[powers - question]

18:13,20 19:2 20:15,22 21:4 21:16,21 22:6 26:17 28:25 29:21,25 30:4 31:7 32:1,19 33:5,12 34:9 34:12 35:12,18 36:2 37:15,22 38:9,15,24 39:10,14,20,23 40:1,4,5,12,17 40:23 41:3,12 41:19 42:3,21 43:1 47:11,18 48:9,12,15 49:15,18,21,22 50:4,9 51:10 51:15,19,23 52:1,3,8,19 53:6,16,18,22 57:20 58:14 59:4,8,20 60:5 60:9,22 61:4,7 61:13 63:1,5 63:15 precipitation 48:23 prep 60:16 preparation 8:8,20 10:8,18 11:12 13:9 15:18 16:6 55:10	prepare 15:6,8 15:9 prepared 58:19 preparing 10:21 13:14 prepping 9:9 presence 54:6 presumed 40:18,24 previously 11:4 11:24 13:3 46:3 primary 18:25 primer 29:14 29:18 30:10,17 30:18 31:12,18 32:5 36:20 38:7 privilege 12:14 privileged 12:22 problem 44:14 problematic 59:1 problems 8:25 9:15,21,25 11:17 14:11,15 15:13,20,22,24 16:1 57:12 procedures 52:10 proceedings 64:7,11,15 proceeds 12:12	process 18:25 21:19,21,23 29:22 43:6 62:20 produce 40:18 produced 32:23 39:4 producing 59:1 product 3:11 3:14,17,19 12:14 30:5,14 30:17 31:8,13 32:2,5,23 36:19 44:20 54:5,13,17 55:2,4 56:10 56:12 productivity 40:13 products 33:2,6 33:9 45:5,9 48:22 50:11 54:16,24,25 55:1,7,9 professional 24:14 professionally 24:19 projects 27:12 prosecution 14:12,14 protocol 43:7 44:24 45:13 protocols 26:21 27:1	provide 10:25 11:6,13,19,23 12:1,23 13:5 21:20,22 49:12 58:22 provided 11:3 13:15 pst 1:20 4:2 public 67:19 publication 58:10,12 publications 23:19 published 50:1 57:3 purification 43:8,22 45:3 purifying 44:9 44:16 50:20 purposes 57:5 pushing 5:17 put 29:13 32:20 51:10 putative 31:21 q quackenbush's 9:16,22 quality 39:5,8 question 6:5,6 6:19 10:5 11:5 11:25 12:16,20 12:23 18:19 19:1 26:17 28:9 29:16,25 35:25 60:23
---	---	---	--

[question - researchers]

62:1 questions 26:13 39:24 45:23,24 46:11 49:16 59:14,15,25 61:1,6,20,22 62:6,8,10,15,19 62:20,23 63:2 quick 48:10 59:4 quite 11:5 25:3 52:10 quote 60:18	40:18,21 41:23 42:2,7,12,15,25 43:9 45:7 46:5 52:15 53:2,12 53:15 59:2,3 61:3 65:9 67:5 reading 35:4 37:17 38:3 reads 38:10 42:1 61:1 really 9:14,19 9:24 11:25 14:10,15 25:24 35:25 reason 65:11 66:6,9,12,15,18 66:21 reasons 37:6 41:12,16 rebuttal 8:3 recall 4:20,21 15:23 16:4,15 27:7 49:11,24 59:15,24 60:25 receipt 65:18 recently 27:23 28:8 recess 34:11 53:21 59:9 61:9 recollection 50:12 recommended 60:20	recommends 51:2 record 48:13 48:14 59:7 61:7 63:15 64:14 recorded 64:12 recross 3:5 61:12 redirect 3:4 59:12 61:15,17 61:25 62:3 reference 56:23 57:4 58:15 referenced 6:25 7:1 65:6 references 14:22 15:2 refresh 50:12 regarded 5:7 regarding 9:20 9:25 16:1 regions 39:5,8 register 12:8 relate 23:22,23 related 19:23 64:17 relevant 11:18 relied 10:17 rely 10:21 remember 50:14 remotely 4:6 removal 59:25 60:3	remove 43:12 45:4,9 48:24 49:6 51:3 removed 39:9 43:6 45:21 56:2 removing 50:24 rendering 9:6 57:9,15 repeat 28:6 29:16 30:11 37:21 38:14 44:11 58:2 62:1 replication 19:11 reported 1:24 reporter 4:7 6:10 19:10 35:9 36:17 47:9 51:6 64:4 64:24,24,25 reporter's 64:1 represent 30:25 31:1,2 31:20 requested 35:9 36:17 47:9 51:6 64:16 required 67:13 requires 52:20 researchers 46:7
r			
r 2:1,12 66:3,3 raised 8:25 raises 57:12 ralph 4:13 rarely 18:6 rather 19:14 43:20 44:25 rcr 3:11,14,17 30:5,14 31:8 32:2,24 reaction 32:24 33:2,7 54:12 54:15 58:19,22 58:23,25,25 reactions 34:22 read 7:24 8:24 11:4,24 37:11 37:14,16 38:1 38:6,13,16,20 38:25 39:3,7 39:15 40:10,15			

[resemble - sequencing]

resemble 29:12 resembles 29:9 reserved 63:17 response 62:19 rest 54:14 resulting 55:3 results 20:17 retain 51:4 retained 47:22 48:5 return 65:13,17 reversible 47:22 review 8:23 9:1 9:3,11,13,16,22 10:3,7,24 14:8 14:10,12,14 15:10,12,17 39:25 57:8,14 64:16 65:7 reviewed 8:19 11:16,17 13:20 13:24 15:11,13 15:15,19 23:19 right 10:13,18 15:5 17:3 19:18 24:15 27:11,19 33:3 34:7,21 35:2 36:24 40:24 41:14 42:4 45:25 46:17 48:19 57:15,24 58:5 63:9	rna 50:24 role 25:2 room 4:15 7:4 45:3 rosati 2:5 rpr 1:24 64:3 64:23 rs 46:4 rules 5:24 running 52:12 s s 2:1 66:3 s2 46:25 sample 52:11 55:24 samples 24:25 47:20 satija 8:14,23 8:24,25 9:2 11:16 57:12 satija's 9:14,18 9:20 10:24 13:20,25 14:11 14:15 15:13,20 15:21,25 save 22:8 saves 21:7,23 22:7,11 saying 23:2 36:12 60:19 says 17:8 25:12 28:7 34:21 35:2 41:20 42:7 43:5,21 45:25 46:18	47:12,17,19 48:21 50:1,3,5 50:8,14,19,19 50:22,23 51:1 51:9 54:4 56:5 56:19 57:2,3 58:18 scenario 23:6 schematic 3:11 3:14,17 30:4 30:13 31:8 32:1 scheme 31:14 school 59:22 scope 17:14 18:5,10,17,23 20:12,19 21:1 21:10 22:3,13 23:1,10 28:5 29:11,23 32:17 33:4 37:19 38:5,12,19 39:6,12,17 40:11,20,25 41:9,15,24 42:14,18,23 47:6,16 51:8 52:7,16 53:4 53:13 seattle 1:22 2:6 4:1 second 9:22 16:16,19 46:13 52:3 58:17	section 43:21 see 5:25 6:22 7:23 17:11 20:9 28:1,18 43:23 46:15 49:1 52:8 54:8 56:25 57:6,23 58:4 seen 13:3 49:22 select 60:20 selected 7:15 10:3,7,10,15,20 47:21 selection 43:22 59:25 60:3 selective 43:8 sense 37:24 sensitivity 53:1 sensors 3:23 sent 65:14 sentence 37:20 46:17 47:12 58:18 september 50:1 sequence 37:16 38:2 42:9 61:3 sequenced 33:10 46:2 sequences 38:10,17 42:11 sequencing 8:9 13:10 17:9,20 18:4,9,12,15,22 19:3,10,21,21 19:24 23:22,23
--	---	---	--

[sequencing - sternekessler.com]

25:1,3,5,13,16 25:18,19,21 26:1,3,9,18 27:22 28:3,9 33:2,7 34:22 35:18 36:3,13 36:18 37:11 38:20 39:1,11 40:7 42:4 46:4 54:12,15 55:18 56:7 60:16 series 19:14 38:6 serve 4:23 served 5:1 24:3 24:5 set 39:7 sg 52:9,24 53:9 sgerrard 2:8 sheared 46:24 47:4,20 55:21 shearing 47:7 sheet 65:11 short 48:25 49:6 53:18 shows 32:23 side 27:19 32:20 34:21 51:10 sign 65:12 signal 41:4,7 signature 63:17 64:23	signed 65:20 similar 13:11 13:12,12 46:3 simply 12:23 single 3:11,14 3:17 27:24 30:5,14 31:8 32:2 33:24 34:1,5,6 42:9,9 46:19 sir 31:15 32:21 53:24 site 29:18 32:15 sites 29:14 30:10,17,19 31:13,18,22 32:5,15 35:22 36:4 sitting 22:19 size 43:8,22 46:24 47:5,21 59:25 60:3,20 sized 55:22 sizes 32:23 47:8 skill 22:20 23:7 51:20 52:25 skilled 17:13,17 17:25 18:13 20:9,17,24 21:8,25 22:10 slab 52:4 smaller 45:5,10 45:13 smrt 27:22 28:3 28:9 35:18	36:2 smrtbell 3:10 3:12,15,18 28:20,21 29:1 29:3,5,8,15,18 30:6,15,23 31:10,11 32:3 32:24 33:24 35:21,24 37:18 38:4 42:10 44:9,16 46:1,7 46:18 54:10,11 54:13,16,24 55:2,2,10,14,17 56:7,13,16 smrtbells 54:19 smws 34:4 solid 47:22 solutions 65:23 sonja 2:5 16:11 sonsini 2:5 sophomores 59:23 sorry 10:2 29:16 30:11 31:10 34:5 37:20 38:13 43:20 44:11 52:2,5 57:25 62:1 sort 39:8 sounds 17:3 speak 14:17 35:9 37:7 47:10 61:15	specific 5:6 62:22,25 specifically 10:9 19:5 38:8 42:1 44:5,6 specifics 21:2 spent 11:2 spin 3:21 48:23 49:3,5,10 50:13,19,20,23 51:3 56:16 spri 47:23 staining 52:13 53:8 stand 44:2 starts 27:20 46:14 52:3 54:2 state 64:24,25 stated 64:9,15 states 1:1 45:2 55:24 56:15 64:5 ste 2:6 stenographic... 64:11 step 43:8 steps 45:3 54:14 stern 4:16 sterne 2:13 4:13 sternekessler.... 2:15,15
--	---	---	--

[stochastic - think]

stochastic 29:22 32:12 store 24:25 strand 42:10 stranded 3:11 3:14,17 30:5 30:14 31:8 32:2 46:18,19 strands 31:2 streptavidin 35:6,8,11,14 36:22 37:1,3,6 strike 31:10 strong 37:5,8,9 strongest 37:3 37:7 structure 29:1 structures 32:11 studies 8:25 subread 39:11 39:13 61:3 subreads 39:16 41:20,23 42:1 42:7,8,17,24 61:1 subscribed 67:14 substance 34:13 61:16,18 63:5,7 suite 1:21 supervision 64:12	supplement 45:21,22,24 supplemental 8:17 45:18 46:11 supply 52:22 supposed 42:3 sure 10:5 28:7 32:8 37:22,23 38:15 40:1 49:18 50:17 53:20 58:1,8 61:18 62:2 suspected 53:11 sworn 4:6 64:7 64:10 67:14 system 27:23 33:23 46:4,25 52:23 58:20 systems 58:21	taken 10:23 34:11 53:21 59:9 61:9 talk 6:10 talks 42:1 target 46:24 teaching 59:21 59:22 team 26:20,25 27:4,6 teams 27:9 tech 50:5,10 51:2 technologies 17:9 19:3 27:24 28:8 technology 8:15 18:4 19:22 28:4,10 28:12,14 47:23 tedious 52:10 tell 22:16 60:15 temperature 45:3 template 3:10 8:8 13:9 28:20 28:21 29:22 32:24 33:24 34:6 35:24 36:4,7,14,19 37:18,18 38:3 38:4,8,11,18 41:6,14 42:10 46:7,19 54:3 54:10 55:2,14	55:17 56:13,16 60:16 templates 35:21 44:9,16 48:24 49:6 56:7 ten 3:19 17:5 32:3 term 37:10,13 38:25 39:10,13 terms 3:20 26:16 40:6 42:6 62:17 63:4 ternary 36:10 36:18,23 testified 4:7 testify 64:10 testimony 9:15 9:20,25 12:12 34:13,17 61:16 61:25 62:3 64:7,11,14 65:9,18 67:8 text 43:5 thank 12:11 31:4 63:14 thing 22:16 things 15:9 17:24 18:6 23:6 24:16 25:17 41:10 think 15:5 22:14 59:5
	t 66:3,3 t4 54:6 tab 8:3,5,6,7,8 8:10,11,12,13 8:14,15,17 table 3:9 7:23 7:25 10:1,2,16 11:7,14,20 12:2,24 13:6 14:5 take 32:19 34:9 48:9 53:16,18 54:1 59:4 60:7		

[third - versus]

third 16:21,23 three 3:12 30:6 30:15 47:19 throughput 19:17,19 time 4:20,21 11:2 16:13,14 16:16,19,21,23 35:16 51:16 52:25 64:9,15 65:19 timeframe 65:8 times 3:13,16 3:19 30:6,15 31:9,11 32:3 today 10:10,11 10:12 16:6 22:19 today's 15:6,18 together 43:17 43:18 told 14:21 took 61:14 top 44:25 topics 61:23,24 62:2,4,12,13,17 topologically 29:3,6 tore 5:18 torn 5:20 total 39:4 55:15 towards 43:5 tpowers 2:15 trademark 1:1	tradeoffs 20:20 traditional 52:24 53:9 trainer 24:10 24:12 training 13:11 transcribed 6:1 64:12 transcript 64:8 64:13,16 65:6 65:20 67:5,8 travers 8:7 13:11 27:14,14 27:14,17 28:17 30:23 32:20 33:1,5,19,19 34:19,20,23 35:1,5,19,20 36:3,13 46:21 48:15,19,21 49:5 53:16,17 53:22 54:1,2,4 56:6,9,15,20,23 57:2,2 58:11 58:12 trey 2:12 trial 1:3 trimming 39:5 true 17:24 36:25 64:14 67:8 truth 64:10 try 6:9 15:1 tuesday 1:20	turn 27:17 two 20:15,23 21:6 22:20 23:3 43:14,17 54:5 61:8 type 63:1,8,9 typical 52:9 u u.s. 3:24 56:23 57:3,8 uh 27:16 ultraviolet 53:8 53:10 um 10:14 26:5 41:21 under 6:14 20:22 21:6 23:6 34:21 48:7 53:8 64:12 undersigned 64:3 understand 6:2 6:4,12,20 10:5 11:10 35:25 understood 6:6 undesirable 55:10,12 united 1:1 university 1:8 4:14 23:25 24:2 65:4 66:1 67:1 unrelated 5:21	usable 41:17 use 25:6 29:15 30:22 35:19,20 36:3 45:12 50:20,23 56:2 56:10 used 9:9 25:8 25:10 34:1 37:6,11 38:25 41:10 46:7 65:20 useful 58:23 user 3:21 uses 34:23 35:5 55:17 using 25:7 32:24 33:1,6 43:7,11 44:8 44:15 46:3,25 48:3,7 49:5 51:3 56:7 usually 53:7 utilizes 27:23 v v 65:4 66:1 67:1 variety 37:6 verbal 6:1 verify 65:9 veritext 65:14 65:23 veritext.com 65:15 versus 22:22
---	---	--	---

[vicky - zoning]

vicky 1:24 64:3 64:23 vii 56:3 visualized 53:7 voluminous 46:12 vs 1:7	24:3,5 26:12 35:9 47:10 48:11 59:5 64:9,20 65:5,8 65:10,12,19 word 46:14 work 11:12 12:14 13:14 14:11,13 24:19 24:19 25:1,3 26:14,23 worked 27:9 50:11 workflow 39:15 41:22 working 11:2 26:25 writing 14:1 wrote 13:23 15:10 wsgr.com 2:7,8 65:2	yom 65:15 york 2:13
w		z
wa 2:6 want 5:23 58:9 59:5 washing 44:21 washington 1:8 1:22,24 2:14 4:1,15 64:6,24 65:4 66:1 67:1 waveguide 34:7 34:8 35:13 36:5,8,15,16 39:5 40:10,18 41:5 waveguides 33:23 35:6 way 28:6 32:14 36:12 42:24 62:5,11,13,20 64:17 we've 34:9 went 55:1 wife 59:21 wilson 2:5,12 withdraw 29:25 witness 2:3 4:6 4:23 5:1 12:6		zero 33:23 34:6 34:8 35:6,13 36:4,8,15,16 39:4 40:10,17 41:4 zmw 36:11,21 36:24 38:8 40:16,22 41:2 41:7,14,18 58:24,25 zoning 44:13
	x	
	x 3:1 xp 47:24 48:1,3	
	y	
	yeah 12:11,13 12:21 38:23 44:12 years 13:11 yesterday 16:25 17:1,2 yield 20:16	

Federal Rules of Civil Procedure

Rule 30

(e) Review By the Witness; Changes.

(1) Review; Statement of Changes. On request by the deponent or a party before the deposition is completed, the deponent must be allowed 30 days after being notified by the officer that the transcript or recording is available in which:

(A) to review the transcript or recording; and

(B) if there are changes in form or substance, to sign a statement listing the changes and the reasons for making them.

(2) Changes Indicated in the Officer's Certificate. The officer must note in the certificate prescribed by Rule 30(f)(1) whether a review was requested and, if so, must attach any changes the deponent makes during the 30-day period.

DISCLAIMER: THE FOREGOING FEDERAL PROCEDURE RULES ARE PROVIDED FOR INFORMATIONAL PURPOSES ONLY.

THE ABOVE RULES ARE CURRENT AS OF APRIL 1, 2019. PLEASE REFER TO THE APPLICABLE FEDERAL RULES OF CIVIL PROCEDURE FOR UP-TO-DATE INFORMATION.

VERITEXT LEGAL SOLUTIONS
COMPANY CERTIFICATE AND DISCLOSURE STATEMENT

Veritext Legal Solutions represents that the foregoing transcript is a true, correct and complete transcript of the colloquies, questions and answers as submitted by the court reporter. Veritext Legal Solutions further represents that the attached exhibits, if any, are true, correct and complete documents as submitted by the court reporter and/or attorneys in relation to this deposition and that the documents were processed in accordance with our litigation support and production standards.

Veritext Legal Solutions is committed to maintaining the confidentiality of client and witness information, in accordance with the regulations promulgated under the Health Insurance Portability and Accountability Act (HIPAA), as amended with respect to protected health information and the Gramm-Leach-Bliley Act, as amended, with respect to Personally Identifiable Information (PII). Physical transcripts and exhibits are managed under strict facility and personnel access controls. Electronic files of documents are stored in encrypted form and are transmitted in an encrypted fashion to authenticated parties who are permitted to access the material. Our data is hosted in a Tier 4 SSAE 16 certified facility.

Veritext Legal Solutions complies with all federal and State regulations with respect to the provision of court reporting services, and maintains its neutrality and independence regardless of relationship or the financial outcome of any litigation. Veritext requires adherence to the foregoing professional and ethical standards from all of its subcontractors in their independent contractor agreements.

Inquiries about Veritext Legal Solutions' confidentiality and security policies and practices should be directed to Veritext's Client Services Associates indicated on the cover of this document or at www.veritext.com.

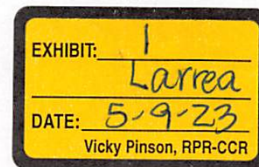


Table of Contents

TAB	DOCUMENT
Tab 1	EX 1033 - Larrea Rebuttle Declaration
Tab 2	EX 1018 - US 20090298075A1
Tab 3	EX 1020 - US 7,476,503
Tab 4	EX 1021 - Travers
Tab 5	EX 1035 - Template Preparation Sequencing Guide
Tab 6	EX 1036 - Chin 2011
Tab 7	EX 2001 - Eid 2009
Tab 8	EX 2002 - Niedringhaus 2010
Tab 9	EX 2006 - Egan 2012
Tab 10	EX 2014 - Satija Declaration
Tab 11	EX 2021 - PacBio Technology Backgrounder (2010)
Tab 12	EX2026 - Eid Suppl Material

EXHIBIT: 2
Larrea
DATE: 5-9-23
Vicky Pinson, RPR-CCR



EXHIBIT: 3
Larrea
DATE: 5-9-23
Vicky Pinson, RPR-CCR





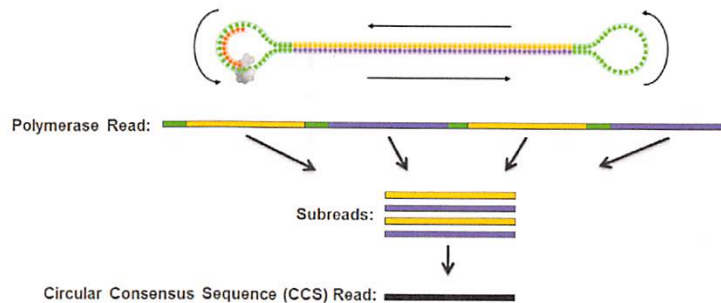
EXHIBIT: 5
Larrea
DATE: 5.9.23
Vicky Pinson, RPR-CCR



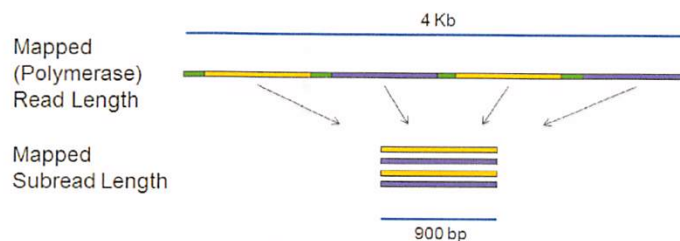
| | | | | | | | | | | | | | | | | | | |

General Technology

- **barcode score:** The alignment score between a read and an ideal barcode sequence. The maximum barcode score is twice the length of the ideal barcode sequence.
- **circular consensus sequencing (CCS) read:** The consensus sequence resulting from alignment between subreads taken from a single ZMW. Note that generation of the CCS read does not include or require alignment against a reference sequence. Generation of CCS reads using the CCS algorithm requires at least two full-pass subreads from the insert.



- **full-pass subread:** A subread that begins at one adapter sequence and ends at another adapter sequence. A full-pass subread does **not** begin or end in the middle of an insert sequence.
- **mapped polymerase read length:** Approximates the sequence produced by a polymerase in a ZMW. The total number of bases along a read from the first adapter of aligned subread to the last adapter or aligned subread.
- **mapped subread length:** The length of the subread alignment to a target reference sequence. This does **not** include the adapter sequence.



- **N50 read length metric:** The read length at which 50% of the bases are in reads longer than, or equal to, this value.
- **paired barcodes:** Barcode sequences that are different (asymmetric) on either end of an insert present in a SMRTbell™ template. The barcoding analysis software uses unique pairs of barcodes to separate and analyze reads.

-
- **polymerase read:** A sequence of nucleotides incorporated by the DNA polymerase while reading a template, such as a circular SMRTbell template. They can include sequences from adapters and from one or multiple passes around a circular template, which includes the insert of interest. Polymerase reads are most useful for quality control of the instrument run. Polymerase read metrics primarily reflect movie length and other run parameters rather than insert size distribution. Polymerase reads are trimmed to include only the high-quality region.
Note: Sample quality is a major factor in polymerase read metrics.
 - **polymerase read length:** The total number of bases produced from a ZMW after trimming the low-quality regions. It may include the adapter sequences.
 - **polymerase read quality:** A trained prediction of a read's mapped accuracy based on its pulse and base file characteristics (peak signal-to-noise ratio, average base QV, inter-pulse duration, and so on).
 - **preassembled long read (PLR):** A generated read that has been output from the preassembly step of HGAP.
 - **productivity:** A measure of the reads from a ZMW. $P=1$ means that there is a polymerase read from that ZMW. $P=0$ means that a ZMW did not produce a read and is presumed to be lacking a polymerase. $P=2$ means "other" and the signal collected from the ZMW was not conducive to efficient base calling, possibly due to multiple template-polymerase complexes bound in the ZMW, high background signal, and so on.
 - **read length:** The number of contiguous bases incorporated into a nascent strand during template-directed synthesis. Read length can be reported as mean, N50, 95th percentile, and max.
 - **read quality (RQ):** The *de novo* prediction of the accuracy of subreads from a single ZMW. Sometimes also referred to as QC Score or Read Score.
 - **SMRT® Sequencing:** The process of nucleic acid sequencing using Pacific Biosciences' single molecule, real-time sequencing technology.
 - **Stage Start:** When specified using RS Remote software, the sequencing reaction starts at the optical stage, just before data acquisition begins. This captures sequence data at the 5' end of the SMRTbell template that would otherwise be missed, and is especially important for long inserts where multi-pass CCS reads will not occur and the longest subreads are critical.
 - **subread:** Each polymerase read is partitioned to form one or more subreads, which contain sequence from a single pass of a polymerase on a single strand of an insert within a SMRTbell template and no adapter sequences. The subreads contain the full set of quality values and kinetic measurements. Subreads are useful for applications such as *de novo* assembly, resequencing, base modification analysis, etc.
 - **symmetric barcodes:** Barcode sequences that are identical on both ends of an insert present in a SMRTbell template.
 - **zero-mode waveguide (ZMW):** A nanophotonic device for confining light to a small observation volume. This can be, for example, a small hole in a conductive layer whose diameter is too small to permit the propagation of light in the wavelength range used for detection. Part of a SMRT Cell.

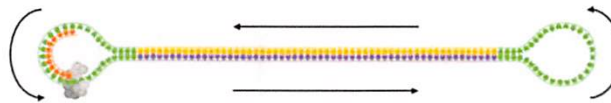
Consumables

- **DNA Controls:** DNA control template-polymerase complexes that are ready for loading on the instrument. There are two types of controls: **Spike-in** controls and **stand-alone** controls. Spike-in controls are added into your samples for sequencing. Stand-alone controls are run on a single SMRT Cell for instrument troubleshooting purposes.
- **MagBead:** Small paramagnetic bead, 2-3 μm in size. The DNA-polymerase complexes are attached to the magnetic beads, which can then be pulled down for easy removal of contaminants from the supernatant during the binding step. The DNA-polymerase complex/bead mixture can be used for the on-instrument immobilization step. See also **MagBead loading**.
- **SMRT[®] Cells:** Consumable substrates comprising arrays of zero-mode waveguide nanostructures. SMRT Cells are used in conjunction with the **DNA Sequencing Kit** for on-instrument DNA sequencing.

Template Preparation

- **AT ligation:** The library construction protocol option by which an adapter with a single-nucleotide T overhang is ligated to an insert with a single-nucleotide A overhang. The workflow that uses this ligation option also contains an A-tailing step.
- **barcode padding:** An optional 5 base pair constant sequence appended to unique barcode sequences. Can be used to normalize ligation of adapters during template preparation.
- **barcoded adapter:** A SMRTbell adapter with a barcode sequence appended to the end of the stem region. When using barcoded adapters, SMRTbell templates will have a symmetric barcode structure.
- **barcoded SMRTbell[™] template:** A SMRTbell template with two barcoded adapters.
- **blunt ligation:** The library construction protocol option by which an adapter lacking any overhangs is ligated to an insert also lacking any overhangs. The workflow that uses this ligation option also lacks the A-tailing step.
- **diffusion loading:** Immobilization of DNA-polymerase complex into the ZMWs on the SMRT Cell via diffusion. Smaller inserts load preferentially compared to larger inserts.
- **DNA damage repair:** A step in the SMRTbell template preparation that repairs a variety of types of DNA damage, including pyrimidine dimers, abasic sites, and nicks.
- **DNA end repair:** A step in the SMRTbell template preparation that removes 5' and 3' overhangs, and phosphorylates 5' ends.
- **DNA fragmentation:** The generation of smaller DNA fragments. Multiple methods may be used to fragment DNA, including hydrodynamic shearing, mechanical shearing, sonication, and enzymatic digestion.
- **input DNA:** Purified DNA used as an input to the **DNA Template Prep Kit**.
- **insert size:** The length of the double-stranded nucleic acid fragment in a SMRTbell template, excluding the hairpin adapters.

- **MagBead loading:** Immobilization of large DNA molecules into the ZMWs on the SMRT Cell chip via MagBeads. The smallest inserts, hairpin dimers, and excess polymerase are washed out in the initial MagBead binding and washing steps. As a result, medium and larger size inserts load better and have a higher sequencing accuracy (compared to diffusion loading of similar-sized inserts).
- **polymerase binding:** The binding of the sequencing polymerase to an appropriate binding site on a nucleic acid template.
- **primed template:** Refers to a template molecule that is annealed with primer. A primed template is a product of the template prep protocol and an input to the binding protocol.
- **primer annealing:** The hybridization of a sequencing primer to an appropriate binding site on a template.
- **size selection:** The removal of unwanted fragments from a mixture based on size. This can refer to the removal of only the shortest fragments, such as adapter dimers, or to the isolation of a very narrow range of insert sizes. Depending on the size range of interest and the equipment available, size selection can be accomplished with AMPure PB beads, manual isolation from an Agarose-gel, automated gel isolation, or BluePippin™ or SageELF™ size-selection systems.
- **SMRTbell™ template:** A double-stranded DNA template capped by hairpin adapters (i.e., SMRTbell adapters) at both ends. A SMRTbell template is topologically circular and structurally linear, and is the library format created by the **DNA Template Prep Kit**.



- **template:** Molecules to be sequenced; the **DNA Template Prep Kit** produces SMRTbell templates.
- **template annealing:** Process of hybridizing primer(s) to nucleic acid templates.
- **template library:** A set of nucleic acid molecules to be sequenced; the **DNA Template Prep Kit** process generates template libraries.
- **template-polymerase complex:** Primed template bound to DNA polymerase; the output of the **DNA/Polymerase Binding Kit** process.

General Software

- **PacBio® DevNet** (<http://www.pacb.com/products-and-services/analytical-software/devnet/>): Resource for informatics researchers, and life scientists; includes data sets, source code, application programming interfaces, and documentation.
- **SMRT® Link**: Web-based end-to-end workflow manager for the Sequel™ System. It includes software applications for setting up samples, designing and monitoring sequencing runs, and analyzing and managing sequence data.
 - **Data Management**: SMRT Link software module used to create Projects and Data Sets, and manage access permissions for Projects and users.
 - **SMRT® Analysis**: SMRT Link software module used to perform secondary analysis of data including sequence alignment, variant detection, *de novo* assembly, RNA analysis, and detecting epigenetic DNA modifications.
- **SMRT® View**: Java-based genome browser used to visualize aligned or assembled reads. Part of the SMRT Analysis suite.

Sequel System only:

- **Sequel Instrument Control Software (ICS)**: Software that controls instrument functionality including: on-instrument operations, user interface, and primary analysis.
- **Run Design**: SMRT Link software module used to design and import sequencing runs into Sequel ICS.
- **Run QC**: SMRT Link software module that displays quality metrics for individual instruments and runs on the Sequel System.
- **Sample Setup**: SMRT Link software module used to calculate binding and annealing reactions for preparing DNA samples for use on the Sequel System.

PacBio RS II System only:

- **barcode FASTA**: A FASTA-format file used by barcoding software to identify ideal barcode sequences. For *symmetric* barcodes, each barcode sequence identifies a single bin for demultiplexing reads. For *paired* barcodes, each unique pair of barcodes should be listed as two sequentially-named FASTA sequences.
- **Binding Calculator**: Web-based application used to calculate binding and annealing reactions for preparing DNA samples for use on the PacBio RS II.
- **RS Dashboard**: Web-based software that displays quality metrics for individual instruments, runs, and SMRT Cells.
- **RS Remote**: Windows-based client software used to design and monitor sequencing runs. Directs user to look at primary analysis data in-depth using RS Dashboard.
- **RS Touch**: Touchscreen user interface on the instruments to help the user load the instrument and start a run. Provides direct feedback on instrument status.
- **SMRT™ Pipe**: Command-line interface of SMRT Analysis suite. SMRT Pipe is used to launch secondary analysis jobs.
- **SMRT™ Portal**: Web-based software used to help set up secondary analysis jobs and view quality reports. Part of the SMRT Analysis suite.

Primary Analysis

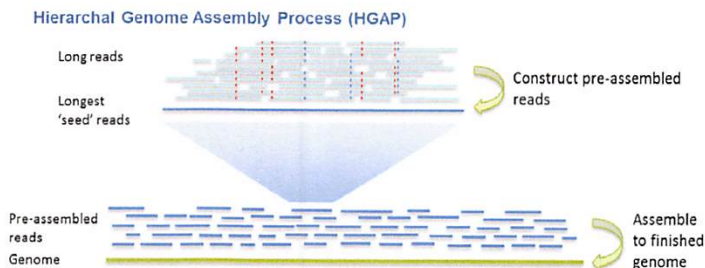
- **high-quality (HQ) region:** Annotates the high-quality sequencing regions of a read to be used during raw read trimming.
- **movie time:** The time specified for collecting data from a SMRT Cell.
- **primary analysis:** On-instrument analysis which includes signal processing of the movie, base calling of the traces and pulses, and quality assessment of the base calls. Subsequently, it trims the sequences to the high-quality (HQ) regions, identifies adapter, barcode (optional), and control sequence reads, assigns read scores, and outputs the subread data in a BAM file.
- **pulse:** The representation of an incorporation event derived from a trace. It includes metrics such as interpulse duration, pulse height, and pulse width.
- **QV Metric:** Phred-like scores that reflect the probability of a correct call for each base.
- **raw read trimming:** Removing a section of the raw read such as a low-quality region. Trimming of an unfiltered read produces a polymerase read.
- **reads/SMRT® Cell:** The number of reads generated per SMRT Cell.

Secondary Analysis

- **Arrow:** (Note: This replaces Quiver.) A highly accurate consensus and variant caller that can generate >99.999% accurate consensus sequences using local realignment and the full range of quality scores associated with reads. Part of SMRT Analysis.
- **barcode FASTA:** A FASTA-format file used by barcoding software to identify ideal barcode sequences. For symmetric barcodes, each barcode sequence identifies a single bin for demultiplexing reads. For paired barcodes, each unique pair of barcodes should be listed as two sequentially-named FASTA sequences.
- **BLASR:** Used for targeted sequencing. Maps reads against a reference; part of SMRT Analysis.
- **consensus accuracy:** Accuracy based on aligning multiple sequencing reads or subreads together.
- **circular consensus accuracy:** Accuracy based on consensus sequence from multiple sequencing passes around a single circular template molecule.
- **circular consensus sequence analysis:** Processing of sequencing data generated by circular consensus sequencing to create a circular consensus read.
- **circular consensus sequencing (CCS):** Sequencing performed on a circular template in which a subread was generated during a sequencing pass around the template. These reads are aligned to each other to generate a single high-accuracy consensus read. CCS analysis to generate consensus reads requires CCS data with at least two full-pass subreads.
- **consensus sequence determination:** Generation of a consensus sequence from multiple individual reads or from multiple subreads of the same DNA molecule (CCS). Also termed "consensus calling". Multi-molecule consensus for targeted applications can be generated using

the Long Amplicon Analysis (LAA) protocol. For other multi-molecule applications, such as whole-genome sequencing and *de novo* assembly, consensus is generated using the Arrow consensus caller as part of the resequencing protocol.

- **HGAP:** The Hierarchical Genome Assembly Process (HGAP) is an algorithmic approach for generating high-quality *de novo* assemblies using PacBio data. HGAP includes pre-assembly, *de novo* assembly and assembly polishing steps. HGAP3 utilizes an AssembleUnitig module for *de novo* assembly and Quiver for polishing. HGAP4 uses Falcon for *de novo* assembly and Arrow for polishing.



- **Quiver:** (Note: This is replaced by Arrow.) A highly accurate consensus and variant caller that can generate >99.999% accurate consensus sequences using local realignment and the full range of quality scores associated with reads. Part of SMRT Analysis.
- **secondary analysis:** Follows primary analysis and uses basecalled data. It is application-specific, and may include:
 - Filtering/selection of data that meets a desired criteria (such as quality, read length, etc.).
 - Comparison of reads to a reference or between each other for mapping and variant calling, consensus sequence determination, alignment and assembly (*de novo* or reference-based), variant identification, etc.
 - Quality evaluations for a sequencing run, consensus sequence, assembly, etc.
 - PacBio's SMRT Analysis contains a variety of secondary analysis applications including RNA and Epigenomics analysis tools.
- **secondary analysis application** (Formerly "Secondary analysis protocol"): A secondary analysis workflow that may include multiple analysis steps. Examples include *de novo* assembly, resequencing, RNA and epigenomics analysis.
- **SMRT[®] Analysis Suite:** Client/server software that performs automated and distributed analysis of sequencing data generated by the PacBio System.
- **tertiary analysis:** Analyses following secondary analysis, which includes comparisons of secondary analysis results across different samples, application-specific analyses, variant classification, and disease/gene annotations.

Base Modification

- **amplified control:** Control created by separately sequencing an amplified version of the sample of interest.
- **interpulse duration (IPD):** Metric for the length of time between emission pulses indicative of base incorporation events. Base modifications in a template molecule can impact IPD, so changes in IPD are used to detect base modifications during SMRT Sequencing.
- **IPD Ratio:** The ratio of the mean IPD of a native sample to the mean IPD in a second sample or in silico control at a position of inquiry in a template.
- **in silico control:** Computational model for predicting the mean IPD per given sequence context at the position of inquiry.

Instrument Terminology

PacBio RS II System only:

- **Blade Center:** The podium containing computers used to process primary analysis and control the instrument.
- **carrier plate:** Metal plate with 12 slots to accommodate 8Pac strips.
- **drawer:** Each of the two compartments where consumables are placed by the user (including the door). The left drawer houses reagents and samples, and the right drawer houses SMRT Cells and tips.
- **Environmental Cabinet:** The right-most compartment of the instrument containing the nitrogen intake and chiller sub-module.
- **insert in drawer:** The removable aluminum assembly found in the (left) reagent drawer. The insert consists of a top unit with septa clamps and tube openings, and a bottom unit with receptacle areas for two reagent plates, one mixing plate, one sample plate, and four tubes. The insert is removable for cleaning.
- **tip station:** The area of the SMRT Cell and tip drawer that accommodates up to 6 boxes of pipette tips.

For Research Use Only. Not for use in diagnostic procedures. © Copyright 2010 - 2016, Pacific Biosciences of California, Inc. All rights reserved. Information in this document is subject to change without notice. Pacific Biosciences assumes no responsibility for any errors or omissions in this document. Certain notices, terms, conditions and/or use restrictions may pertain to your use of Pacific Biosciences products and/or third party products. Please refer to the applicable Pacific Biosciences Terms and Conditions of Sale and to the applicable license terms at <http://www.pacb.com/legal-and-trademarks/product-license-and-use-restrictions/>.

Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT, SMRTbell, Iso-Seq and Sequel are trademarks of Pacific Biosciences. BluePippin and SageELF are trademarks of Sage Science, Inc. NGS-go and NGSengine are trademarks of GenDx. All other trademarks are the sole property of their respective owners. PN 000-710-267-06



CHROMA SPIN™ Columns User Manual

PT1300-1 (PR49719)
Published 09 September 2004

Table of Contents

I. Introduction	3
II. List of Components	7
III. CHROMA SPIN™ Columns Protocol	9
A. General Considerations	9
B. Column Capacity	10
C. Purification using CHROMA SPIN Columns	11
D. Buffer Exchange	12
E. Additional Hints	13
IV. CHROMA SPIN™ 1000 HT 96-Well Plate Protocol	14
V. References	16
VI. Related Products	16
Appendix A: Conversion of RCF to rpm	17
Appendix B: Column Performance	18
List of Tables	
Table I. Buffers and applications	4
Table II. Recommended sample concentration for various types of molecules	11
Table III. Range of sample volumes for CHROMA SPIN Columns	11
Table IV. Size of protein molecules retained by various columns	18
Table V. Size of molecules removed by the column matrix	18
Table VI. Size of molecules recovered in the eluate	19
List of Figures	
Figure 1. Nucleic acid purification using CHROMA SPIN Columns	5
Figure 2. Standard CHROMA SPIN Column and Collection Tubes	9
Figure 3. The percentage removal of dsDNA fragments as a function of size for each of the five pore-size CHROMA SPIN Columns	20

I. Introduction

Spin-column chromatography, which combines economy, effectiveness of gel filtration, and speed of centrifugation, is a common method of choice for nucleic acid purification. **CHROMA SPIN™ Columns** are routinely used to purify DNA or RNA from smaller contaminants such as salts, solvents, nucleotides, enzymes, proteins, polysaccharides, metabolites, and in some cases other (smaller) nucleic acids. In addition, CHROMASPIN columns can be used for size selection of nucleic acids within a wide size range (from 15 nucleotides to >1,000 bp). Nucleic acids purified with these columns are well suited for use in many molecular biology experimental procedures, including sequencing, labeling, PCR, cloning, and in vitro transcription (Sambrook & Russell, 2001).

CHROMASPIN columns are based on the principle of gel filtration chromatography and are packed with resins designed to selectively remove contaminating molecules based on size. The proprietary resins are comprised of uniform microscopic beads of a hydrophilic porous material. Molecules larger than the pore size are excluded from the resin. These molecules quickly move through the gel bed when the column is centrifuged briefly, while molecules smaller than the pore size are held back (Hagel et al., 1989; Porath et al., 1959). Thus, DNA and protein molecules are eluted from the column in order of decreasing molecular size. Highest sample recovery (>90%) is achieved when the molecule of interest is significantly larger than the matrix pore size; highest purity (>90% enrichment) is achieved when contaminants are significantly smaller than the matrix pore size.

These columns are available in six matrix pore sizes that cover a wide range of applications. The number at the end of each column name indicates the approximate 50% retention point. The smallest pore-size column, CHROMA SPIN-10, allows the selective removal of molecules as small as NaCl or dNTPs; CHROMASPIN-1000, the largest pore-size column, will remove 50% of molecules as large as 1,000-bp DNA fragments. In addition, there are four intermediate pore-size columns, CHROMA SPIN-30, -100, -200, and -400, which can be used for the removal of intermediate-size nucleic acids and proteins. There is no defined upper size limit of nucleic acid that can be purified. We have successfully used these columns to purify yeast genomic DNA (~79 kb). Please refer to Appendix B for additional information on purification and removal characteristics of the various pore-size columns.

CHROMA SPIN columns are provided sterilized and ready to use, with a choice of three different column buffers (TE, STE, and DEPC-treated H₂O). In many cases, pre-equilibrating the column before use is not required (Table I); however, the column matrix may be re-equilibrated with any buffer that is appropriate for your application. Nucleic acid purification can be accomplished in just a few steps (Figure 1), and the entire procedure takes only ~9 min, including prespin time. The purified molecules are eluted in sterile column buffer and automatically collected in a self-contained, nuclease-free storage tube.

I. Introduction continued

TABLE I: BUFFERS AND APPLICATIONS

Buffer	Application
TE (10 mM Tris-HCl [pH 8.0], 1.0 mM EDTA)	Routine DNA work
STE (TE buffer + 0.1 mM NaCl)	Purification of plasmid DNA and other applications requiring high salt concentration
DEPC-water (water and matrix treated with 0.1% DEPC, autoclaved)	A nuclease-free buffer for RNA work

There is no need to transfer the eluted nucleic acid sample after purification—this reduces the risk of spills, losses, and accidental exposure in the case of radioactive samples. (Unincorporated radioactive nucleotides are safely retained in the small volume of the column.)

The separation performance of each CHROMASPIN column has been vigorously tested with a wide range of nucleic acid fragments (Figure 3, Appendix B). All components are sterilized including the column, buffer, gel-filtration matrix, and collection tubes. Also, each new lot is tested to verify that it is nuclease-free.

Applications

The following section provides several applications using CHROMA SPIN columns.

Use any CHROMA SPIN™ Column for the following applications:

- Desalting or exchanging nucleic acid/protein solution buffer
- Removing oils and solvents (such as phenol)
- Removing unincorporated nucleotides, whether they are unlabeled or labeled (Radioactively labeled, biotinylated, and fluorescently-labeled nucleotides are all efficiently removed.)
- Removing all other molecules smaller than 5 kDa

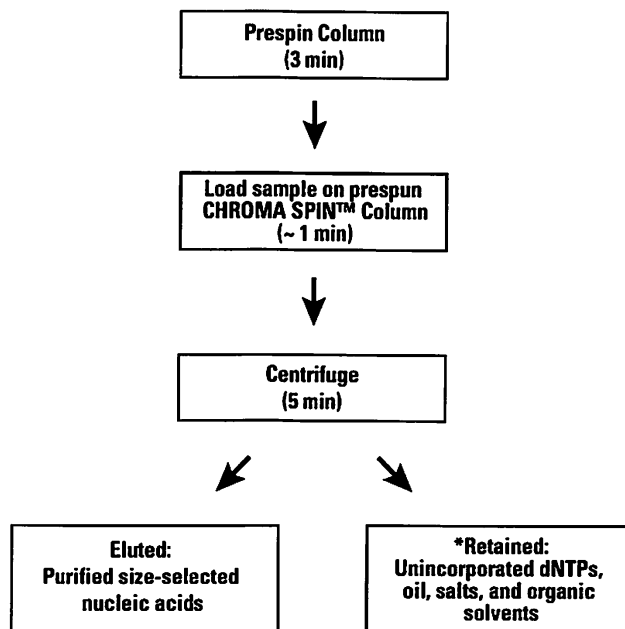
Use CHROMA SPIN™-10 Columns for the following applications:

- Purifying small oligonucleotides (>15 bp) after labeling reactions

Use CHROMA SPIN™-30 Columns for the following applications:

- Purifying DNA and RNA molecules (>35 bp) after labeling reactions
- Purifying proteins >30 kDa
- Removing short primers (<9 bp) after labeling or cDNA synthesis
- Removing unligated DNA linkers (<9 bp) after ligation reactions

I. Introduction continued



Total time: ~ 9 min

*Other types of contaminating molecules depend on their size and the pore-size column used (Appendix B).

Figure 1. Nucleic acid purification using CHROMA SPIN™ Columns.

Use CHROMA SPIN™-100 Columns for the following applications:

- Recovering DNA molecules >140 bp at a high yield
- Purifying DNA molecules >100 bp after PCR amplification (high purity)
- Purifying RNA molecules >140 bp after in vitro transcription
- Eliminating unextended primers, short extension products, and digested DNA fragments <30 bp
- Removing enzymes and proteins <250 kDa
- Removing digested agarose polymers

I. Introduction continued

Use CHROMA SPIN™-200 Columns for the following applications:

- Recovering DNA molecules >350 bp at a high yield
- Purifying DNA molecules >150 bp after PCR amplification (high purity)
- Removing PCR primers <50 bp and primer-dimers <45 bp
- Removing proteins, including enzymes <1,000 kDa

Use CHROMA SPIN™-400 Columns for the following applications:

- Purifying (or size selecting) DNA molecules >600 bp after PCR amplification or cDNA synthesis
- Purifying small plasmids >600 bp
- Removing larger primers, extension products, and digested DNA fragments <100 bp
- Size selecting larger mRNAs by removing short (<100 nucleotides) RNA molecules prior to cDNA synthesis and cloning
- Removing proteins, including enzymes <8,000 kDa

Use CHROMA SPIN™-1000 Columns for the following applications:

- Purifying DNA molecules >1,350 bp (including λ DNA, plasmids, and yeast genomic DNA)
- Size-selecting large cDNAs (>1,350 bp) for library construction
- Removing RNA and DNA molecules <300 bp
- Removing all proteins, including enzymes

Note: For additional information on choosing the correct matrix pore size, see Appendix B.

II. List of Components

Store all CHROMA SPIN products at room temperature (10–32°C). Do not freeze. Columns are packed in sterile, sealed plastic bags.

CHROMA SPIN™+ STE Columns

Cat. Nos. 636055 & 636060

- 20 CHROMA SPIN+STE Columns
Column matrix is equilibrated and stored in sterile STE buffer
(0.1 M NaCl, 10 mM Tris-HCl [pH 8.0], 1 mM EDTA)
- 40 Sterile 2-ml Microcentrifuge Collection Tubes

CHROMA SPIN™+ STE Columns

Cat. Nos. 636056, 636058 & 636061

- 50 CHROMA SPIN+STE Columns
Column matrix has been equilibrated and stored in sterile STE buffer
- 100 Sterile 2-ml Microcentrifuge Collection Tubes

CHROMA SPIN™+ TE Columns

Cat. No. 636072

- 20 CHROMA SPIN+TE Columns
Column matrix is equilibrated and stored in sterile TE buffer
(10 mM Tris-HCl [pH 8.0], 1 mM EDTA)
- 40 Sterile 2-ml Microcentrifuge Collection Tubes

CHROMA SPIN™+ TE Columns

Cat. Nos. 636066, 636069, 636073, 636076, 636079 & 636082

- 50 CHROMA SPIN+TE Columns
Column matrix has been equilibrated and stored in sterile TE buffer
- 100 Sterile 2-ml Microcentrifuge Collection Tubes

CHROMA SPIN™ + DEPC-H₂O Columns

Cat. Nos. 636089, 636093 & 636096

- 20 CHROMA SPIN + DEPC-H₂O Columns
Column matrix has been equilibrated and stored in H₂O +
0.1 mM EDTA (pH 8.0) treated with diethyl-pyrocabonate (DEPC),
and then autoclaved.
- 40 Sterile 2-ml Microcentrifuge Collection Tubes

II. List of Components continued

CHROMA SPIN™ + DEPC-H₂O Columns

Cat. Nos. 636087, 636090 & 636094

- 50 CHROMA SPIN + DEPC-H₂O Columns
Column matrix has been equilibrated and stored in H₂O +
0.1 mM EDTA (pH 8.0) treated with diethyl-pyrocabonate (DEPC),
and then autoclaved
- 100 Sterile 2-ml Microcentrifuge Collection Tubes

CHROMA SPIN™ 1000 HT 96-Well Plate

Cat. Nos. 636251

- CHROMA SPIN-1000 Plate (including top and bottom seals)
Column matrix has been equilibrated and stored in H₂O +
0.1 mM EDTA (pH 8.0) treated with diethyl-pyrocabonate (DEPC).
- Collection Plate

III. CHROMA SPIN™ Column Protocol

A. General Considerations

- If you wish to replace the column storage buffer with a different buffer, see Section III.D.
- We recommend centrifuging CHROMA SPIN Columns in a swinging bucket or horizontal rotor to allow the sample to pass through the column matrix uniformly. Fixed-angle rotors can also be used, but there is a risk that a portion of the sample will slide down the inner side of the column instead of passing through the gel matrix, resulting in reduced or inconsistent purification.
- To determine the rpm appropriate for your centrifuge rotor, see Appendix A. Alternatively, a clinical (tabletop) centrifuge can be used.

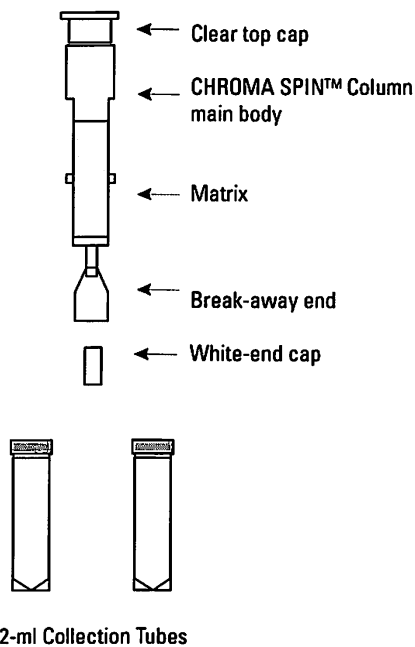


Figure 2. Standard CHROMA SPIN™ Column and Collection Tubes.

III. CHROMA SPIN™ Column Protocol continued

B. Column Capacity

1. Sample concentration

Typical molecular biology applications involve concentrations of nucleic acids that are well below the saturation limit for the resin in CHROMA SPIN Columns. Nevertheless, there is a recommended upper limit on the concentration of protein or nucleic acid in the sample to be purified (Table II). The upper concentration limits are determined by three factors: (1) the size of the molecules, (2) the potential for random aggregation, and (3) the viscosity of the sample. If the concentration exceeds the recommended limits indicated in Table II, there is a risk that the molecules (a) will not be completely dissolved, (b) will form aggregates, or (c) will form a highly viscous solution. Any one of these factors or a combination thereof will reduce the effectiveness of the column by impeding the progress of large molecules through the gel bed and/or preventing small particles from entering the matrix pores.

Notes:

- If your sample is above the recommended maximum concentration and you notice a problem with viscosity, aggregation, or incomplete dissolution, then dilute the sample.
- If the DNA or RNA in the sample to be purified is below the recommended minimum concentration, add carrier (e.g., sonicated calf thymus DNA, glycogen, or tRNA) to a final concentration of 0.02–0.1 mg/ml. This will improve recovery of the desired molecules from the column.

2. Sample volume

The volume of sample that a CHROMA SPIN Column can accommodate depends on the type (pore size) of column matrix (Table III). It is important to not exceed the maximum recommended sample volume for each column; excessive sample volume may cause small molecules, normally retained in the matrix pores, to be eluted. Each column also requires that a minimum sample volume be loaded for efficient recovery of purified molecules. If the sample volume is below the recommended minimum, there is a risk that some of the desired large molecules will not be efficiently recovered.

Notes:

- If the sample volume is less than the recommended minimum volume, dilute the sample before loading it on the column — or load the undiluted sample onto the prespun column and apply an additional aliquot of buffer to make up the difference between the sample volume and the recommended minimum volume.
- Please see Appendix B for additional information on how matrix pore size influences column performance.

III. CHROMA SPIN™ Column Protocol continued

TABLE II: RECOMMENDED SAMPLE CONCENTRATIONS FOR VARIOUS TYPES OF MOLECULES

Molecules in the sample	Optimum concentration range
NTPs, dNTPs	up to 4 mM each
DNA (up to 50 kb)	0.02–1.0 mg/ml
DNA or RNA (<100 bp)	0.02–1.5 mg/ml
Protein (>30,000 Da)	<500 µM (<20 mg/ml)

TABLE III: RANGE OF SAMPLE VOLUMES FOR CHROMA SPIN™ COLUMNS

Column (standard capacity)	Sample volume
CHROMA SPIN™-10, -30	25–50 µl
CHROMA SPIN™-100, -200	40–75 µl
CHROMA SPIN™-400, -1000	70–100 µl

C. Purification using CHROMA SPIN™ Columns

1. Upon removing a CHROMA SPIN Column from the protective plastic bag, invert it several times to resuspend the gel matrix completely.
2. Holding the CHROMA SPIN Column upright, grasp the break-away end between your thumb and index finger and snap off (Figure 2). Place the end of the spin column into one of the 2-ml microcentrifuge (collection) tubes provided, and lift off the top cap. Save the top cap and the white-end cap.
3. Centrifuge at 700 x g for 5 min. After centrifugation, the column matrix will appear semi-dry. This step purges the equilibration buffer from the column and reestablishes the matrix bed.
4. Remove the spin column and collection tube from the centrifuge rotor, and discard the collection tube and column equilibration buffer.
Note: Always hold both the CHROMA SPIN Column and the 2-ml microcentrifuge tube when removing them from the rotor.
5. Place the spin column into the second 2-ml microcentrifuge tube. Carefully and slowly apply the sample to the **center** of the gel bed's flat surface. Do not allow any sample to flow along the inner wall of the column. (Refer to Table III for recommended sample volumes.)

III. CHROMA SPIN™ Column Protocol continued

Note: For particularly small sample volumes (e.g., <30 µl), a conventional, tapered 1.5-ml microcentrifuge tube can be substituted for the 2-ml collection tube. This will allow the sample to be confined to a narrower area for easier handling.

6. Centrifuge at 700 x g for 5 min.
7. Remove the spin column and collection tube from the rotor and detach them from each other. The purified sample is at the bottom of the collection tube.

Note: Hold the sample collection tube to prevent it from detaching from the spin column. This is particularly important when radioactive samples are processed. For radioactive samples, seal the used spin column with the top and end caps prior to transfer and disposal.

D. Buffer Exchange

For those applications in which none of our column storage buffers are convenient for storing the purified nucleic acid sample, the column matrix may be re-equilibrated with a more appropriate buffer for the particular application.

1. Invert the CHROMA SPIN Column several times to completely resuspend the gel matrix.
2. Remove the top cap first and then the bottom cap from the column. Save the caps. Place the bottom tip of the column gently (snugly, but not tightly) into one of the 2-ml microcentrifuge tubes provided.
3. Discard any buffer that immediately collects in the 2-ml microcentrifuge tube. Replace the column gently into the same microcentrifuge tube.
4. Centrifuge the column in a swinging bucket rotor or in a fixed-angle rotor at 700 x g for 3 min.
5. Discard the collected buffer from the 2-ml tube. Replace the column gently in the tube. Add 1 ml of the new buffer to the gel. After the first application of the new buffer, replace both caps on the column and resuspend the matrix completely by inverting the column several times.
6. Let the buffer drain through the column by gravity or centrifuge again at 700 x g for 3 min.
7. Empty the 2-ml collection tube and replace the column gently in the same 2-ml tube. Repeat the buffer application and centrifugation twice more.

Note: After the first application and removal of buffer, do not invert the column or resuspend the matrix because disrupting the matrix bed will reduce the final yield.

8. If the column is going to be used right away, remove the buffer according to Section III.C, Steps 3 & 4, and proceed with purification (Section III.C.5–7).

Note: If the column will be saved for later use, close the column with the bottom cap and add 0.7 ml of the new buffer. Replace the top cap tightly, and invert the column several times to resuspend the gel matrix. Store the column at 4°C or room temperature (22°C) in an upright position. **Do not freeze.**

III. CHROMA SPIN™ Column Protocol continued

E. Additional Hints

- The column should be in an upright position at all times (except when centrifuging) to avoid sample back-flow.
- The column fitted in the 2-ml microcentrifuge tube can be placed inside a 17 x 100-mm polypropylene tube during centrifugation in a swinging bucket rotor for ease of handling.
- For easier sample loading on the gel matrix, add Bromophenol Blue (0.01% w/v) to the sample. The dye molecules will be removed by the column, along with other small molecules.
- Do not overload or underload the column. For optimal column performance, follow the recommended sample volumes, as described in Section III.B.
- To remove small primers from larger DNA products or templates that contain sequences complementary to the primers, heat-denature the sample before loading it on the column. This will reduce hybridization between primers and larger DNAs and result in better separation.
- For best separation results, keep the sample concentration within the recommended range described in Section III.B. If necessary, dilute the sample or add carrier.

IV. CHROMA SPIN™ 1000 HT 96-Well Plate Protocol

PLEASE READ ENTIRE PROTOCOL BEFORE STARTING.

This protocol is recommended for purifying sample volumes of 50–100 µl/well. For best results, DNA or RNA (up to 50 kb) should have a concentration between 0.02–1.0 mg/ml, and proteins (> 30 kDa) should have a concentration less than 20 mg/ml.

A. Plate Capacity

1. Sample concentration

Many molecular biology applications involve concentrations of nucleic acids that are well below the saturation limit for the resin in the CHROMA SPIN Plate. Nevertheless, there is a recommended upper limit on the concentration of protein or nucleic acid in the sample to be purified (Table II). The upper concentration limits are determined by three factors: (1) the size of the molecules, (2) the potential for random aggregation, and (3) the viscosity of the sample. If the concentration exceeds the recommended limits indicated in Table II, there is a risk that the molecules (a) will not be completely dissolved, (b) will form aggregates, or (c) will form a highly viscous solution. Any one of these factors or a combination thereof will reduce the effectiveness of the column by impeding the progress of large molecules through the gel bed and/or preventing small particles from entering the matrix pores.

Notes:

- If your sample is above the recommended maximum concentration and you notice a problem with viscosity, aggregation, or incomplete dissolution, then dilute the sample.
- If the DNA or RNA in the sample to be purified is below the recommended minimum concentration, add carrier (e.g., sonicated calf thymus DNA, glycogen, or tRNA) to a final concentration of 0.02–0.1 mg/ml. This will improve recovery of the desired molecules from the column.

2. Sample volume

The CHROMASPIN-1000 Plate is designed to accommodate samples of 50–100 µl/well. It is important not to exceed the maximum recommended sample volume for each column; excessive sample volume may cause small molecules, normally retained in the matrix pores, to be eluted. A minimum sample volume should be loaded for efficient recovery of purified molecules. If the sample volume is below the recommended minimum, there is a risk that some of the desired large molecules will not be efficiently recovered.

Note: If the sample volume is less than the recommended minimum volume, dilute the sample before loading it on the column — or load the undiluted sample onto the prespun column and apply an additional aliquot of buffer to make up the difference between the sample volume and the recommended minimum volume.

IV. CHROMA SPIN™ 1000 HT 96-Well Plate Protocol cont.

B. Purification using the CHROMA SPIN™ HT 96-Well Plate

1. Remove the plate from the vacuum-sealed pouch, and then resuspend the gel matrix by inverting the plate several times.
2. Hold the plate upright, and carefully remove the top and bottom seals. Place the 96-Well Plate on the provided Collection Plate.
3. Purge the equilibration buffer and re-establish the matrix bed by placing the assembly (the 96-Well Plate on top of the Collection Plate) in a centrifuge rotor that is designed for 96-well plates, and centrifuging it at 700 x g for 5 min. After centrifuging, the matrix will be semi-dry.
4. While taking care to hold both parts of the assembly together, remove the assembly from the rotor and discard the equilibration buffer from the Collection Plate.
5. Replace the 96-Well Plate on top of the Collection Plate. Carefully and gradually apply each sample to the center of the top surface of the gel bed. Do not add the sample so quickly that it flows around the gel bed and along the inner wall of the column.
6. Once all of the samples are loaded, centrifuge the assembly at 700 x g for 5 min.
7. Remove the assembly from the centrifuge, and carefully detach the Collection Plate, now containing the purified samples, from the 96-Well Plate.

Note: Samples can be stored in the Collection Plate.

V. References

- Hagel, L., Lundstrom, H., Andersson, T. and Lindblom, H. (1989) Properties, in theory and practice, of novel gel filtration media for standard liquid chromatography. *J. Chromatography* 476:329–344.
- Porath, J. and Flodin, P. (1959) Gel filtration: a method for desalting and group separation. *Nature* 183:1657–1659
- Sambrook, J. and Russell, D.W. (2001) *Molecular Cloning: A Laboratory Manual*, (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY).

VI. Related Products

Product	Qty.	Cat. No.
CHROMA SPIN™+ STE Columns		
• CHROMA SPIN+STE-10 Columns	20	636055
	50	636056
• CHROMA SPIN+STE-30 Columns	50	636058
• CHROMA SPIN+STE-100 Columns	20	636060
	50	636061
CHROMA SPIN™+TE Columns		
• CHROMA SPIN+TE-10 Columns	50	636066
• CHROMA SPIN+TE-30 Columns	50	636069
• CHROMA SPIN+TE-100 Columns	20	636072
	50	636073
• CHROMA SPIN+TE-200 Columns	50	636082
• CHROMA SPIN+TE-400 Columns	50	636076
• CHROMA SPIN+TE-1000 Columns	50	636079
CHROMA SPIN™+DEPC-H₂O Products		
• CHROMA SPIN-30+DEPC-H ₂ O Columns	50	636087
• CHROMA SPIN-100+DEPC-H ₂ O Columns	20	636089
	50	636090
• CHROMA SPIN-200+DEPC-H ₂ O Columns	20	636096
• CHROMA SPIN-1000+DEPC-H ₂ O Columns	20	636093
	50	636094
• CHROMA SPIN-1000 HT 96-Well Plate	1	636251

Appendix A: Conversion of RCF to rpm

Revolutions per minute (rpm) calculation for different rotor sizes:

$$\text{rpm} = 1000 \times \sqrt{\frac{\text{RCF}}{1.12 \times r}}$$

or

$$\text{RCF} = \left(\frac{\text{rpm}}{1000} \right)^2 \times 1.12 \times r$$

RCF = relative centrifugal field (g Force = RCF x g)

r = radius (in mm)

Examples:

- a. Swinging bucket rotor: Beckman Model TJ-6, r = 125 mm
rpm = **2236** for RCF = 700 (i.e., 700 x g)
- b. Fixed-angle rotor: Eppendorf Model 5415 C, r = 50 mm
rpm = **3536** for RCF = 700 (i.e., 700 x g)

Appendix B: Column Performance

For each type of column, the size of the matrix bed, the sample volume, and the conditions of centrifugation (speed and time) have been optimized for maximum purification and minimum loss of the desired sizes of nucleic acids. Table IV lists the cutoff sizes of protein molecules that are removed by CHROMA SPIN-10, -30, -100, -200, and -400 Columns at 90% efficiency (no data is available on the upper limit of the size of proteins purified using CHROMA SPIN-1000). Table V lists cutoff sizes of nucleic acids that are removed by the various columns at efficiencies of 90% and 99%. Table VI lists the cutoff sizes of nucleic acid molecules that are recovered in the eluate at efficiencies of 50, 70, and 90%. Figure 3 shows the overall retention and recovery characteristics of the various columns.

TABLE IV: SIZE OF PROTEIN MOLECULES RETAINED BY VARIOUS COLUMNS

Column	Estimated MW of proteins removed (daltons); greater than 90% removal
CHROMA SPIN™-10	<5,000
CHROMA SPIN™-30	<30,000
CHROMA SPIN™-100	<250,000
CHROMA SPIN™-200	<1 x 10 ⁶
CHROMA SPIN™-400	<8 x 10 ⁶

TABLE V: SIZE OF NUCLEIC ACID MOLECULES REMOVED BY THE COLUMN MATRIX

Removal: percent of input molecules removed from the sample		
Column	>90%	>99%
CHROMA SPIN™-10	<4 bp (NTPs, dNTPs)	NaCl
CHROMA SPIN™-30	<17 bp	<9 bp
CHROMA SPIN™-100	<50 bp	<30 bp
CHROMA SPIN™-200	<70 bp	<50 bp
CHROMA SPIN™-400	<170 bp	<100 bp
CHROMA SPIN™-1000	<420 bp	<300 bp

Appendix B: Column Performance continued

TABLE VI: SIZE OF NUCLEIC ACID MOLECULES RECOVERED IN THE ELUATE

Column	Recovery percent of input molecules recovered in the eluate		
	>50%	>70%	>90%
CHROMA SPIN™-10	10 bp	>15 bp	>30 bp
CHROMA SPIN™-30	30 bp	>35 bp	>75 bp
CHROMA SPIN™-100	100 bp	>140 bp	>230 bp
CHROMA SPIN™-200	200 bp	>300 bp	>450 bp
CHROMA SPIN™-400	400 bp	>600 bp	>950 bp
CHROMA SPIN™-1000	1,000 bp	>1,350 bp	>2,000 bp

Appendix B: Column Performance continued

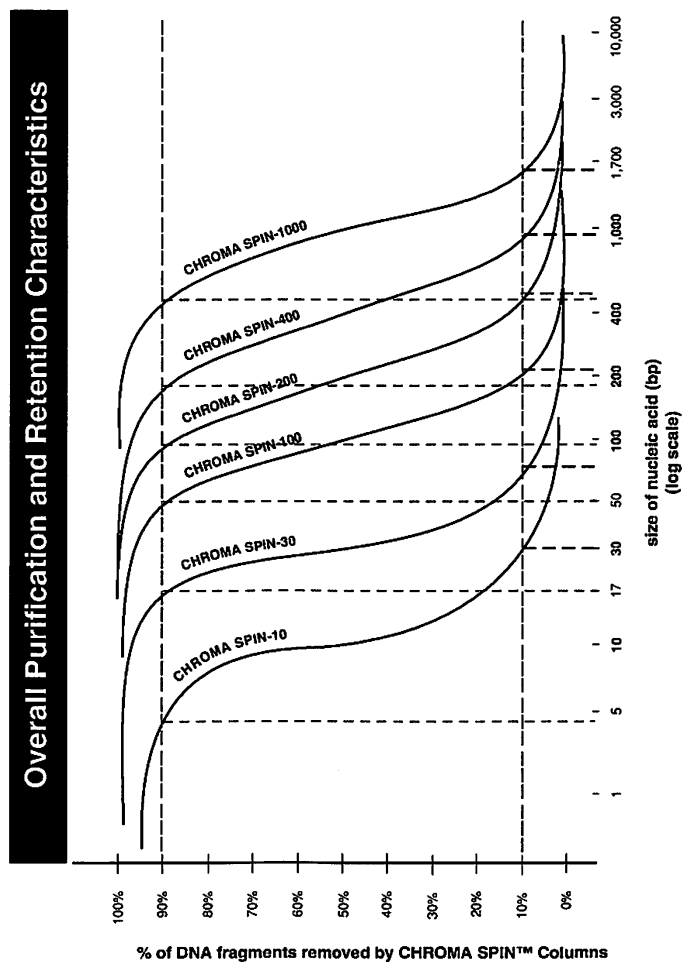


Figure 3. The percentage removal of dsDNA fragments as a function of size for each of the five pore-size CHROMA SPIN™ Columns. Sample volumes used to generate the curves were in the middle of the recommended ranges.

Notes

Notes

Notes

Notice to Purchaser

This product is intended to be used for research purposes only. It is not to be used for drug or diagnostic purposes, nor is it intended for human use. Clontech products may not be resold, modified for resale, or used to manufacture commercial products without written approval of Clontech Laboratories, Inc.

Clontech, Clontech logo and all other trademarks are the property of Clontech Laboratories, Inc. Clontech is a Takara Bio Company. ©2005



Contents lists available at ScienceDirect

Sensors and Actuators B: Chemical

journal homepage: www.elsevier.com/locate/snb

Short communication

Miniaturized gel electrophoresis system for fast separation of nucleic acids

Zhenqing Li^a, Jiaxin Huang^a, Bo Yang^a, Qingxiang You^a, Shinichi Sekine^b, Dawei Zhang^{a,*}, Yoshinori Yamaguchi^{c,d,**}^a Engineering Research Center of Optical Instrument and System, Ministry of Education, Shanghai Key Lab of Modern Optical System, University of Shanghai for Science and Technology, No. 516 JunGong Road, Shanghai 200093, China^b Department of Preventive Dentistry, Graduate School of Dentistry, Osaka University, Yamadaoka Suita-city, Osaka, 565-0871, Japan^c Photonics and Bio-medical Research Institute, Department of Physics Faculty of Science, East China University of Science and Technology, No. 130 Meilong Road, Shanghai, 200237, China^d Department of Applied Physics, Graduate School of Engineering, Osaka University, Yamadaoka Suita-city, Osaka, 565-0871, Japan

ARTICLE INFO

Article history:

Received 17 April 2017

Received in revised form 6 July 2017

Accepted 12 July 2017

Available online 14 July 2017

Keywords:

Slab gel electrophoresis

Capillary electrophoresis

Polymerase chain reaction

Nucleic acids

ABSTRACT

Slab gel electrophoresis (SGE) is very common tool for DNA, RNA and protein analysis, but it is tedious, labor-intensive, skill-dependent, and relatively slow. Herein, we developed a biochip based SGE system, which can resolve the DNA fragments and record their separation process. By electrophoresis of 50 bp DNA ladder, we found that the 16 DNA fragments were effectively resolved within 14 min. In order to validate its feasibility and practicability, we take periodontal pathogens (e.g., *Porphyromonas gingivalis* (P.g), *Tannerella forsythia* (T.f), and *Treponema denticola* (T.d)) as an example by separating their polymerase chain reaction products. Experiments demonstrated that P.g, T.d and T.f were diagnosed within 12 min, and the electrophoresis of P.g showed that the detection limit of this system was about 6.4 ng/μL. Such a low cost system is easy to operate, and can effectively improve SGE in the biological experiment, especially for the labs in the third world countries.

© 2017 Published by Elsevier B.V.

1. Introduction

Nucleic acid separation is an important technology in molecular biology, following by polymerase chain reaction (PCR) analysis or before cloning, sequencing, northern or southern blotting. So far, high performance liquid chromatography (HPLC) [1], capillary electrophoresis (CE) [2], and slab gel electrophoresis (SGE) [3] are three common methods used for the separation of DNA, although research about the new methods are still under way [4–6]. In HPLC, the separation mechanism involves partitioning of analytes between a mobile and stationary phase. However, besides the separation performance, electrophoresis are deemed to have distinct advantages over HPLC, because instrumentation of HPLC is rather complex compared to the simplicity of CE or SGE system. Furthermore, the elution of large molecules can only be achieved by changing the selectivity during separation process (e.g. gradi-

ent elution), and expensive prepacked columns are required which have limited life time in HPLC [7].

Compared with HPLC, the separation channel in CE is an open or gel-filled tube, which was first developed by Mikkers et al. for separation of small anions [8]. It becomes an attractive analysis method in biochemistry field due to its numerous advantages [9,10], such as less reagent consumption, short running times, and high resolution. For example, Luraschi et al. described the use of CE coupled with immunosubtraction for the detection and characterization of low concentration of γ heavy chains in serum [11]. Godelieve Mariën attempted to detect and characterize the μ heavy chain disease by CE [12]. Our group has also applied CE for the detection of periodontal pathogens and *Escherichia coli* [13,14]. With the development of microfabrication technology, research on the CE chip became increasing popular in the past two decades. However, fabrication of the chip was based on expensive photolithography or tedious chemical etching techniques [15]. Walczak's group studied the separation and extraction of cell-free fetal DNA in a CE chip, which was etched in glass [16]. They also developed a miniaturized instrument based on this electrophoresis chip with laser induced fluorescence technology [17]. Although separation of single-stranded and double-stranded DNA was achieved in a

* Corresponding author.

** Corresponding author at: Department of Applied Physics, Graduate School of Engineering, Osaka University, Yamadaoka Suita-city, Osaka, 565-0871, Japan.

E-mail addresses: dzwzhang@usst.edu.cn (D. Zhang),yoshi.yamaguchi@ap.eng.osaka-u.ac.jp (Y. Yamaguchi).<http://dx.doi.org/10.1016/j.snb.2017.07.064>

0925-4005/© 2017 Published by Elsevier B.V.

Table 1
Primer sequence for periodontal pathogens.

Target Gene (16S rRNA gene)		Primers (5'-3')	Product size (bp)
Porphyromonas gingivalis	FwRv	CTGATCTGCAGCAGGTAGCCTC GCACATAACCGGTAGCCTCTAC	132
Tannerella forsythia	FwRv	GCCTATGTAACCTGCCCGCA TGCTTCAGTGTCTATACCT	641
Treponema denticola	FwRv	GAGTAGTCCGTCGGAGTCTATG AATCTCGTATCCGTTCCG	364

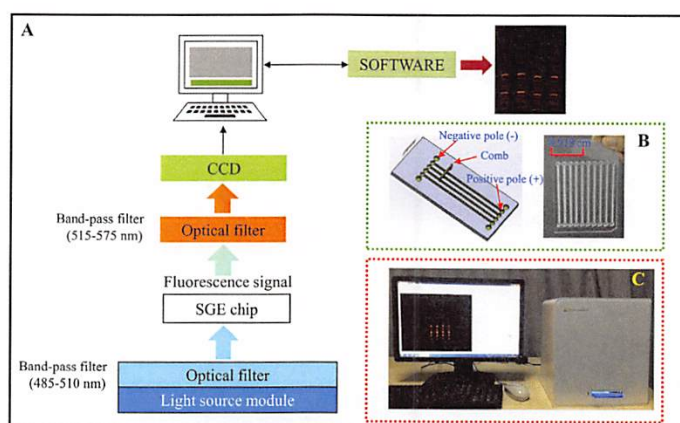


Fig. 1. The miniaturized SGE system: (A) Schematic diagram. (B) The real SGE instrument. (C) The SGE chip.

few minutes, the throughput is not high since only one sample can be separated in one experiment. In order to make it more ease of operation, fully automated and higher throughput commercial instruments for CE or microchip CE (e.g., Bioanalyzer 2100, 5100 ALP, SV110, MCE-2010) are available for routine clinical testing [18], but the high cost of such instrument hindered its prevalence, especially for the labs in the third world countries.

Slab gel electrophoresis (SGE) is the most common approach for separation and analysis of macromolecules (DNA [19], RNA [20] and proteins [21,22]) because of its low cost and high reproducibility. However, typical SGE experiment procedures are quite tedious, which involves preparing the agarose gel (30~50 min), sample loading (1~5 min), running electrophoresis (30~60 min), gel staining (15~30 min), and imaging DNA bands (1~5 min) [19]. Consequently, it requires quite a lot of instruments such as microwave to dissolve the agarose, 2-D electrophoresis cell and electrophoresis power supply for electrophoresis, gel-imaging system to check the DNA bands in the gel, and hence traditional SGE is labor-intensive, time-consuming, skill-dependent, and poor sensitivity.

Furthermore, the DNA molecules are usually visualized under ultraviolet light by staining with appropriated dyes such as ethidium bromide (EtBr) in traditional SGE experiment, but both ultraviolet and EtBr are suspected to be carcinogen. In order to overcome those drawbacks and increase the versatility of SGE, we developed an integrated miniaturized system which can separate the DNA fragments and record the DNA separation process during electrophoresis. Moreover, compared with the conventional SGE device, the system we developed can greatly save the experiment time (about 10 min) and take only 1/20 chemical reagents. Meanwhile, such a chip is reusable and the system is easy to operate. Finally, we take three harmful periodontal pathogens (e.g., *Por-*

phyromonas gingivalis (P.g), *Tannerella forsythia* (T.f), and *Treponema denticola* (T.d)) [23] as an example to validate our system.

2. Materials and methods

2.1. Materials and reagents

SpeedSTAR HS DNA Polymerase, 50bp, 100bp DNA ladder and 10 000× SYBR Green I were from Takara (Shanghai, China). 10× TBE (1× TBE = 89 mM Tris/89 mM boric acid/2 mM EDTA, pH=8.4) buffer and agarose powder were bought from Solarbio (Beijing, China). 1.0% agarose was obtained by diluting 0.1 g agarose into 10 ml 0.5× TBE. The primers for the periodontal pathogens were synthesized by Sangon Biotech (Shanghai, China). Hydroxyethyl cellulose (HEC, 1300k) was from Polysciences (Warrington, PA, USA).

2.2. Capillary electrophoresis

The detailed information of the self-built CE was described in Ref [24]. The electric field is supplied by a high-voltage power supply (MODEL 610E, TREK, Medina, NY, USA). Total capillary length (ID/OD = 75/365 μm) is 6 cm (4 cm effective length), and the capillary is covalently coated with polyacrylamide [25,26]. The excitation wavelength from a mercury lamp is filtered to be 460–495 nm by the optical filter (U-MWIB-3, Olympus, Japan). The fluorescence signal from the conjugate of SYBR Green I and the nucleic acid is detected by a R928 photomultiplier tube (Hamamatsu Photonics, Japan). Power supply and data collection are controlled by LabVIEW (National Instrument, USA). The CE system is placed in a dark box. DNA sample is electrokinetically introduced into the capillary (100 V/cm, 1.0 s). The capillary is flushed by ster-

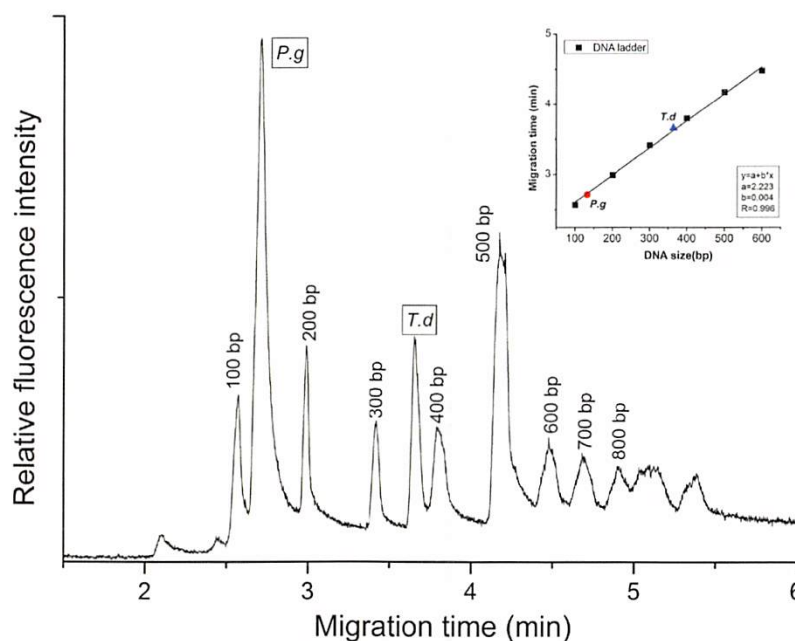


Fig. 2. Separation of mixture of PCR products (P.g and T.d) and 50bp DNA ladder markers by CE. Electrophoresis conditions, polymers: 0.5% HEC (1300k); sample loadings: 100 V/cm (1.0 s); total length and effective length of the capillary 6 cm/4 cm; electric field strength: 100 V/cm.

ilized water after each run. All separations are performed at 26 °C in the air conditioner controlled clean room.

2.3. PCR protocols

PCR was performed in the T100™ Thermal Cycler (Bio-Rad, USA) and programmed with an initial step of denaturation at 95 °C for 2 min. The cycling conditions described as follows: denaturation (95 °C, 10 s) and annealing (64 °C, 30 s). The reaction was performed with 0.7 μl template and 34.3 μl reaction volumes containing 3.5 μl 10 × Fast Buffer I, 2.8 μl dNTP mixture (2.5 μM), 0.7 μl primer 1 (Fw) and 0.7 μl primer 2 (Rv), 0.175 μl Speed-STAR HS DNA Polymerase, and 26.425 μl pure water. The primers for the target gene were tabulated in Table 1. The PCR product was directly analyzed for CE without any further sample preparation.

2.4. Preparation of SGE chip

The blueprint of the SGE chip was drawn in CoreDRAW. The SGE chip (Fig. 1B) in a polymethylmethacrylate (PMMA) block consists of two cylindrical holes for the positive and negative electrode, and arrays of wells for the sample loading, which are made by a comb after the agarose gel is formed in the channel. The chip was fabricated by a laser-engraving machine. For the SGE chip, the surface of the bottom would be quite rough if the engraving mode of the laser-engraving machine is applied, so we used the cutting mode instead. Afterward the SGE chip was successively cleaned by ethyl alcohol and pure water. Finally, we tape a PCR film (Thermo Scientific, Ref:125529) at the bottom of the SGE chip. The geometry parameters of the channel depend on the specific DNA fragments to be resolved. For example, the optimal parameters for the channel are 2.5 × 4.0 × 90 mm if 100 bp DNA ladders are separated.

2.5. SGE system construction

The schematic of the SGE instrument is depicted in Fig. 1A, which mainly consists of the light source, two band-pass optical filters, the SGE chip, CCD, and the software. Photograph of the assembled miniaturized SGE instrument is shown in Fig. 1C. The excitation wavelength from the light source is filtered to be 485–510 nm first, which corresponds to central excitation wavelength of the conjugate of SYBR Green I and single-stranded DNA. The emission wavelength is collected by CCD after the fluorescence filtered by another 515–575 nm band-pass filter. The electric field strength and data procession are controlled by a homemade software.

3. Results and discussion SYBR Green I

3.1. Evaluation of PCR primer selectivity and specificity

Firstly, we evaluated the size of PCR products of P.g and T.d by CE, since the primers of T.f has already proved in our previous study [13]. Each target genes were amplified 35 cycles according to the PCR protocol, then the PCR products were diluted 10 times prior to CE, and finally the mixture of PCR products and 100 bp DNA ladder was resolved in 0.5 × TBE buffer containing 0.5% HEC (1300k) and 1 × SYBR Green I. CE was carried out at 100 V/cm and the results was depicted in Fig. 2. It showed that the mixture was well resolved within 6 min. By negative control PCR (without DNA template) of T.d and T.f, we found that only the peaks corresponding to the primers were observed in the electropherogram (data not shown). The size of PCR products was calculated by linear fitting of DNA size and its migration time, which was described in the insert of Fig. 2. It showed that DNA fragment between 100 bp and 600 bp was linearly related to the migration time, and the correlation coefficient R was 0.996, and thus the size of the PCR products of P.g (132 bp) and T.d (364 bp) was determined.

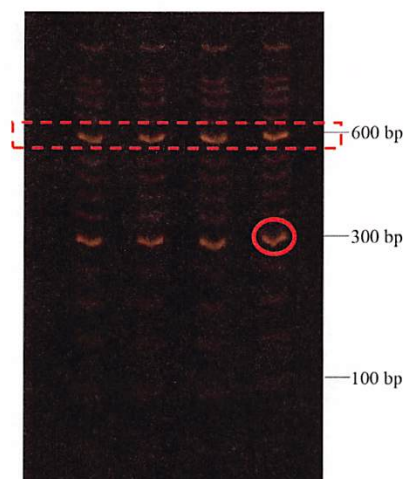


Fig. 3. Separation of 50 bp DNA ladder in the self-developed SGE system. The geometry parameter of the channel is $4.0 \times 4.0 \times 90$ mm.

3.2. Detection of periodontal pathogens in the miniaturized SGE system

Then, we carried out the electrophoresis of *P.g.*, *T.d.* and *T.f.* in the miniaturized SGE system. Prior to SGE, we have performed PCR for each periodontal pathogen with 35 cycles in the T100™ Thermal Cycler according to the PCR protocol. After PCR, the concentration of each periodontal pathogen amplicon was quantified using the Qubit® 3.0 Fluorometer (Q33216, Invitrogen, USA), and it showed that the concentration for *P.g.*, *T.d.* and *T.f.* was 228.67 ng/μl, 175.67 ng/μl and 189.67 ng/μl, respectively. The precast gel in the SGE chip was $0.5 \times$ TBE containing 1.0% agarose. We found that the electric current in the channel increases linearly with the width of the channel in the chip, thus it leads the increase of joule heating in the gel. For example, the DNA looks bent in the gel (Fig. 3) when 50 bp DNA ladders are separated in the $4.0 \times 4.0 \times 90$ mm SGE chip. However, each DNA band moved the same distance in each channel when the electrophoretic conditions are identical, indicating the

high reproducibility of our method. Moreover, it is hard to introduce the sample into the gel if the width of the chip is too narrow. Therefore, the optimal dimension of the channel in the SGE chip we used in the instrument is $2.5 \times 4.0 \times 90$ mm, and the electric field strength enforced along the channel was 20 V/cm. The procedure to make the precast gel in the chip is depicted in Ref [27], and the preparing process was also displayed in the video (See the supporting movie).

Furthermore, we separated the 50 bp DNA ladder in this instrument, which consists of 10 fragments between 50 and 500 bp in multiples of 50 bp and additional fragments at 600, 700, 800, 900, 1000 and 1500 bp. According to the manual, the 300 and 600 bp fragments are brighter than the other fragments and serve as visible reference indicator, and other fragments appear with equal intensity on the gel. Before SGE, the DNA sample was mixed with SYBR Green I, and then 4 μl mixture was added into the precast gel in each channel of the chip. Experiment demonstrated that the current in the agarose is 8.2 mA when the electric field strength is 20 V/cm. Results (Fig. 4A) showed that nearly all DNA fragments could be resolved when the electrophoresis ran about 12 min. The migration distance of each DNA fragment was evaluated by calculating the pixel of DNA band from the well in the chip. We found that DNA migration distance in the agarose gel was linearly related to the logarithm of DNA size (Fig. 4B) ($R=0.980$), indicating that DNA bands in the chip can be employed to estimate the size of the unknown DNA.

3.3. Limit of detection

Finally, we performed the electrophoresis of 50 bp DNA ladder and PCR products of *P.g.* with different amplification cycles ($5 \times$, $15 \times$, $25 \times$ and $35 \times$) in the chip. The electric field strength along the channel was 20 V/cm, and the concentration of the agarose was 1.0%. The electrophoresis ran about 14 min (Fig. 5A). It shows that the resolution of 50 bp DNA ladder looks better than the resolution in Fig. 3. This is because the migration time of DNA is longer, and the relative distance between the adjacent DNA fragments is larger. For the PCR products of *P.g.*, the DNA bands cannot be discerned if the amplification cycle is $5 \times$. It is possibly because the amplification time is so short that there is no PCR products. When the amplification cycle is $15 \times$, the DNA bands still cannot be discerned, which may be attributed to the concentration of PCR products is lower than the detection limit of this system. For the PCR products of *P.g.*

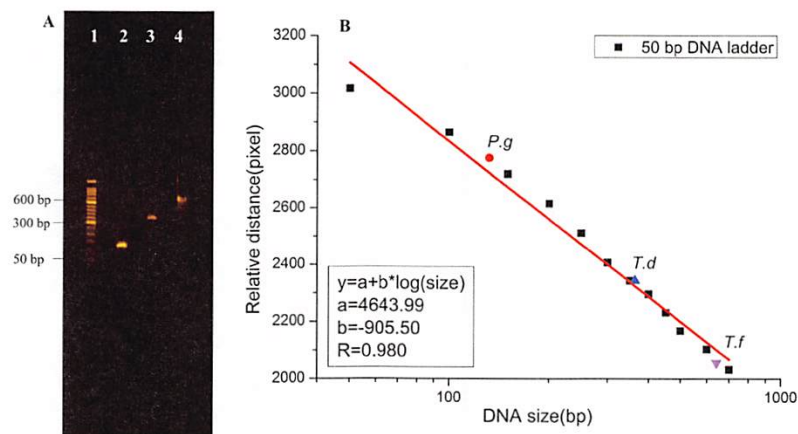


Fig. 4. (A) Separation of 50 bp DNA ladder (lane 1) and PCR products of *P.g.* (lane 2), *T.d.* (lane 3), and *T.f.* (lane 4) in the self-developed SGE system. (B) The dependence of DNA migration distance on the molecular size.

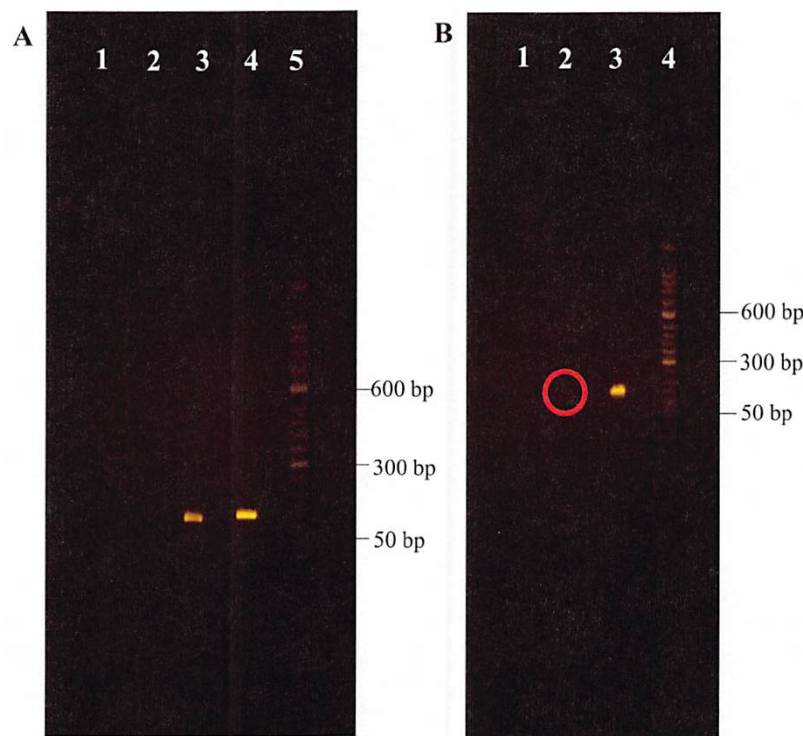


Fig. 5. (A) Separation of 50 bp DNA ladder (lane 5) and PCR products of *P.g* with different amplification cycles: 35 × (lane 4), 25 × (lane 3), 15 × (lane 2) and 5 × (lane 1). (B) Separation of 50 bp DNA ladder (lane 4) and PCR products of *P.g* without dilution (lane 3), diluted by 50 times (lane 2), and 100 times (lane 1).

with 25 × and 35 ×, they migrated nearly the same distance, moreover, the DNA band of 35 × (246 ng/μl) looks brighter than that of 25 × (118.75 ng/μl).

Finally, we have performed the electrophoresis of PCR products of *P.g* diluted by 50, and 100 times. The electrophoretic conditions were the same as above. After 12 min, 50 bp DNA ladders were resolved, which was marked by red circle in Fig. 5B. It showed that the PCR product of *P.g* was observable when it was diluted by 50 times, and the concentration measured by Qubit® 3.0 Fluorometer was about 6.4 ng/μl.

4. Conclusions

We demonstrated an innovative SGE system based on a chip in this work. Such a system cannot only resolve the DNA fragments, but also record the DNA separation process in the gel. Compared with traditional SGE experiment, it demonstrated attractive advantages, such as short run times (<12 min) and low chemical consumption and high band resolution. Furthermore, the low cost SGE system in this work is easy to assemble, which can applied to the estimation of the size of DNA molecules following restriction enzyme digestion, analysis of PCR products, separation of restricted genomic DNA prior to Southern transfer, or RNA prior to Northern transfer, and thus it will greatly improve the efficiency in the biological experiment.

Conflict of interest

The authors have declared no conflict of interest.

Acknowledgements

We gratefully acknowledge support from National Natural Science Foundation of China (No. 61378060 and No. 21205078), Research Fund for the Doctoral Program of Higher Education of China (No.20123120110002) and financial support from University of Shanghai for Science and Technology (No.2017KJFZ049). This work was partially supported by the National key research and development program of China (2016YFB1102303), the National Basic Research Program of China (973Program) (2015CB352001), and National Natural Science Foundation of China (61378060).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.snb.2017.07.064>.

References

- [1] J. Tamaoka, K. Komagata, Determination of DNA base composition by reversed-phase high-performance liquid chromatography, *FEMS Microbiol. Lett.* 25 (1984) 125–128.
- [2] Z. Li, X. Dou, Y. Ni, Y. Yamaguchi, Separation of long DNA fragments by inversion field capillary electrophoresis, *Anal. Bioanal. Chem.* 401 (2011) 1665–1671.
- [3] H. Blum, H. Beier, H.J. Gross, Improved silver staining of plant proteins, RNA and DNA in polyacrylamide gels, *Electrophoresis* 8 (1987) 93–99.
- [4] P. Miao, Y. Tang, B. Wang, F. Meng, Near-infrared Ag2S quantum dots-based DNA logic gate platform for miRNA diagnostics, *Anal. Chem.* 88 (2016) 7567–7573.
- [5] P. Miao, Y. Tang, J. Yin, MicroRNA detection based on analyte triggered nanoparticle localization on a tetrahedral DNA modified electrode followed

- by hybridization chain reaction dual amplification, *Chem. Commun.* 51 (2015) 15629–15632.
- [6] P. Miao, B. Wang, X. Chen, X. Li, Y. Tang, Tetrahedral DNA nanostructure-based microRNA biosensor coupled with catalytic recycling of the analyte, *ACS Appl. Mater. Interfaces* 7 (2015) 6238–6243.
 - [7] J.W. Jorgenson, K.D. Lukacs, Capillary zone electrophoresis, *Science* 222 (1983) 266–272.
 - [8] F.E. Mikkers, F.M. Everaerts, T.P. Verheggen, High-performance zone electrophoresis, *J. Chromatogr. A* 169 (1979) 11–20.
 - [9] F. Liu, J. Wang, L. Yang, L. Liu, S. Ding, M. Fu, et al., Developing a fluorescence-coupled capillary electrophoresis based method to probe interactions between QDs and colorectal cancer targeting peptides, *Electrophoresis* (2016) 2170–2174.
 - [10] J. Wang, C. Zhang, L. Liu, K.A. Kalesh, L. Qiu, S. Ding, et al., A capillary electrophoresis method to explore the self-assembly of a novel polypeptide ligand with quantum dots, *Electrophoresis* 37 (2016) 2156–2162.
 - [11] P. Luraschi, I. Infusino, I. Zorzi, G. Merlini, C. Fundaro, C. Franzini, Heavy chain disease can be detected by capillary zone electrophoresis, *Clin. Chem.* 51 (2005) 247–249.
 - [12] G. Mariën, G. Verhoef, X. Bossuyt, Detection of heavy chain disease by capillary zone electrophoresis, *Clin. Chem.* 51 (2005) 1302–1303.
 - [13] Z. Li, S. Chen, C. Liu, D. Zhang, X. Dou, Y. Yamaguchi, Quantification of periodontal pathogens cell counts by capillary electrophoresis, *J. Chromatogr. A* 1361 (2014) 286–290.
 - [14] Z. Li, D. Li, D. Zhang, Y. Yamaguchi, Determination and quantification of *Escherichia coli* by capillary electrophoresis, *Analyst* 139 (2014) 6113–6117.
 - [15] R. Walczak, W. Kubicki, J. Dziuban, Low cost fluorescence detection using a CCD array and image processing for on-chip gel electrophoresis, *Sens. Actuators B Chem.* 240 (2017) 46–54.
 - [16] W. Kubicki, R. Walczak, Preliminary studies on cell-free fetal DNA separation and extraction in glass lab-on-a-chip for capillary gel electrophoresis, *Procedia Eng.* 47 (2012) 1315–1318.
 - [17] W. Kubicki, R. Walczak, J.A. Dziuban, Miniature instrument for lab-on-a-chip capillary gel electrophoresis of DNA utilizing temperature control technique, *Procedia Eng.* 25 (2011) 1237–1240.
 - [18] S.F. Li, L.J. Kricka, Clinical analysis by microchip capillary electrophoresis, *Clin. Chem.* 52 (2006) 37–45.
 - [19] P.Y. Lee, J. Costumbrado, C.Y. Hsu, Y.H. Kim, Agarose gel electrophoresis for the separation of DNA fragments, *J. Vis. Exp.* 62 (2012) e3923.
 - [20] J.R. Brody, E.S. Calhoun, E. Gallmeier, T.D. Creavalle, S.E. Kern, Ultra-fast high-resolution agarose electrophoresis of DNA and RNA using low-molarity conductive media, *Biotechniques* 37 (4) (2004) 600–602, 598.
 - [21] S. Bhilocha, R. Amin, M. Pandya, H. Yuan, M. Tank, J. LoBello, et al., Agarose and polyacrylamide gel electrophoresis methods for molecular mass analysis of 5- to 500-kDa hyaluronan, *Anal. Biochem.* 417 (2011) 41–49.
 - [22] J.A. Ream, L.K. Lewis, K.A. Lewis, Rapid agarose gel electrophoretic mobility shift assay for quantitating protein: RNA interactions, *Anal. Biochem.* 511 (2016) 36–41.
 - [23] P.J. Ezzo, C.W. Cutler, Microorganisms as risk indicators for periodontal disease, *Periodontol.* 2000 32 (2003) 24–35.
 - [24] Z. Li, Y. Zhao, D. Zhang, S. Zhuang, Y. Yamaguchi, The development of a portable buoyancy-driven PCR system and its evaluation by capillary electrophoresis, *Sens. Actuators B Chem.* 230 (2016) 779–784.
 - [25] S. Hjertén, High-performance electrophoresis: elimination of electroendosmosis and solute adsorption, *J. Chromatogr. A* 347 (1985) 191–198.
 - [26] D. Schmalzing, C.A. Piggee, F. Foret, E. Carrilho, B.L. Karger, Characterization and performance of a neutral hydrophilic coating for the capillary electrophoretic separation of biopolymers, *J. Chromatogr. A* 652 (1993) 149–159.
 - [27] C. Tao, B. Yang, Z. Li, D. Zhang, Y. Yamaguchi, Real-time tracking DNA fragments separation by smartphone, *J. Vis. Exp.* (124) (2017) e55926.

Biographies

Zhenqing Li received his PhD from Shanghai Jiao Tong University. Currently, He is an associate professor of school of optical-electrical and computer engineering of University of Shanghai for Science and Technology. His research interests are capillary electrophoresis and polymerase chain reaction technology.

Jiaxin Huang is a master candidate in University of Shanghai for Science and Technology. She received her bachelor degree in Wuhan Textile University in 2015. Her major research interest covers microfluidic technology.

Bo Yang is a master candidate in University of Shanghai for Science and Technology. He received his bachelor degree in University of Shanghai for Science and Technology in 2015. His major research interest covers microfluidic technology, separation science and technology.

Qingxiang You is a doctor candidate in University of Shanghai for Science and Technology. He received his bachelor degree in Information Engineering University in 2006. His major research interest covers microfluidic technology, separation science and technology and software development.

Shinichi Sekine received his doctor in periodontology from Osaka University in 2003. He is now an associate professor in Department of Preventive Dentistry and Graduate School of Dentistry of Osaka University. His current interests include halitosis and secretory IgA.

Dawei Zhang is a professor in University of Shanghai for Science and Technology. He received his PhD in Shanghai Institute of Optics and Fine Mechanics of Chinese Academy of Sciences in 2005. His research interests are fluorescence detection technology and optic-fluidics.

Yoshinori Yamaguchi received his M.S. and Ph.D. degrees in analytical chemistry from the University of Michigan in 2000 and 2003, respectively. In 2003, he joined the National Institute of Advanced Science and Technology (AIST) as a post-doctor working on the detection of single molecules by surface enhanced Raman Spectroscopy (SERS). From 2004 to 2009, he worked at the Consolidated Research Institute for Advanced Science and Medical Care of Waseda University (ASMeW). He then joined the Photonics Advanced Research Center (PARC) of Osaka University in 2009 where he works as an associate professor. His research interests include single cell analysis, capillary polymer electrophoresis, Raman spectroscopy, and microfluidic devices for point-of-care (POC) biosensors.



US 20090024331A1

(19) **United States**(12) **Patent Application Publication**
Tomaney et al.(10) **Pub. No.: US 2009/0024331 A1**(43) **Pub. Date: Jan. 22, 2009**(54) **METHODS AND PROCESSES FOR CALLING
BASES IN SEQUENCE BY INCORPORATION
METHODS**

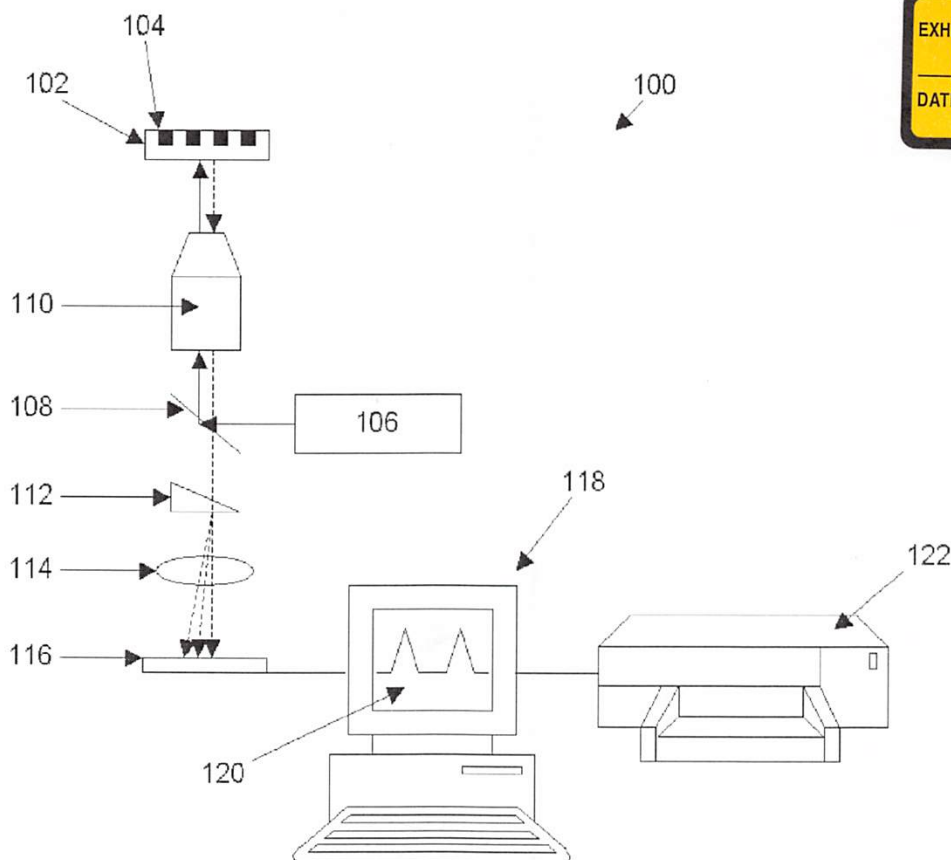
(21) Appl. No.: 12/134,186

(22) Filed: Jun. 5, 2008

(75) Inventors: **Austin B. Tomaney**, San Francisco,
CA (US); **Kenneth Mark
Maxham**, Redwood City, CA (US);
David Holden, Burlingame, CA
(US); **Kevin Hester**, Belmont, CA
(US); **Devon Murphy**, Mountain
View, CA (US); **Patrick Marks**,
San Francisco, CA (US)**Related U.S. Application Data**(60) Provisional application No. 60/933,399, filed on Jun.
6, 2007.**Publication Classification**(51) **Int. Cl.**
G01N 33/68 (2006.01)
G01J 3/28 (2006.01)(52) **U.S. Cl.** 702/19; 356/326(57) **ABSTRACT**

Computer implemented methods, and systems performing such methods for processing signal data from analytical operations and systems, and particularly in processing signal data from sequence-by-incorporation processes to identify nucleotide sequences of template nucleic acids and larger nucleic acid molecules, e.g., genomes or fragments thereof.

Correspondence Address:

**QUINE INTELLECTUAL PROPERTY LAW
GROUP, P.C.
P O BOX 458
ALAMEDA, CA 94501 (US)**(73) Assignee: **Pacific Biosciences of California,
Inc.**, Menlo Park, CA (US)

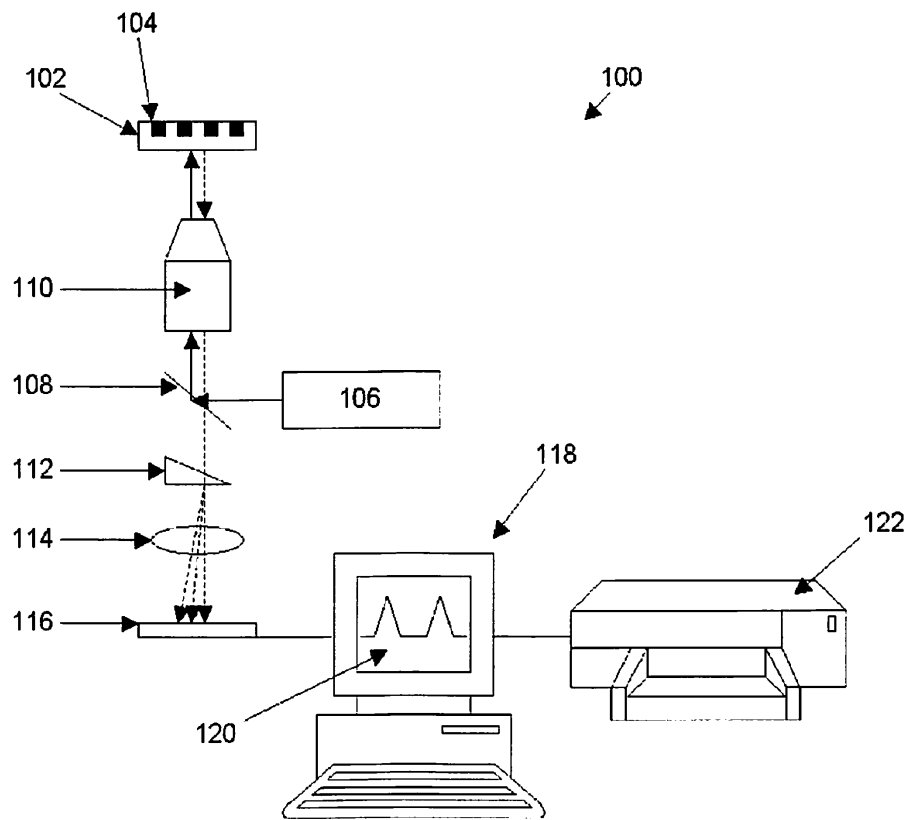


FIGURE 1

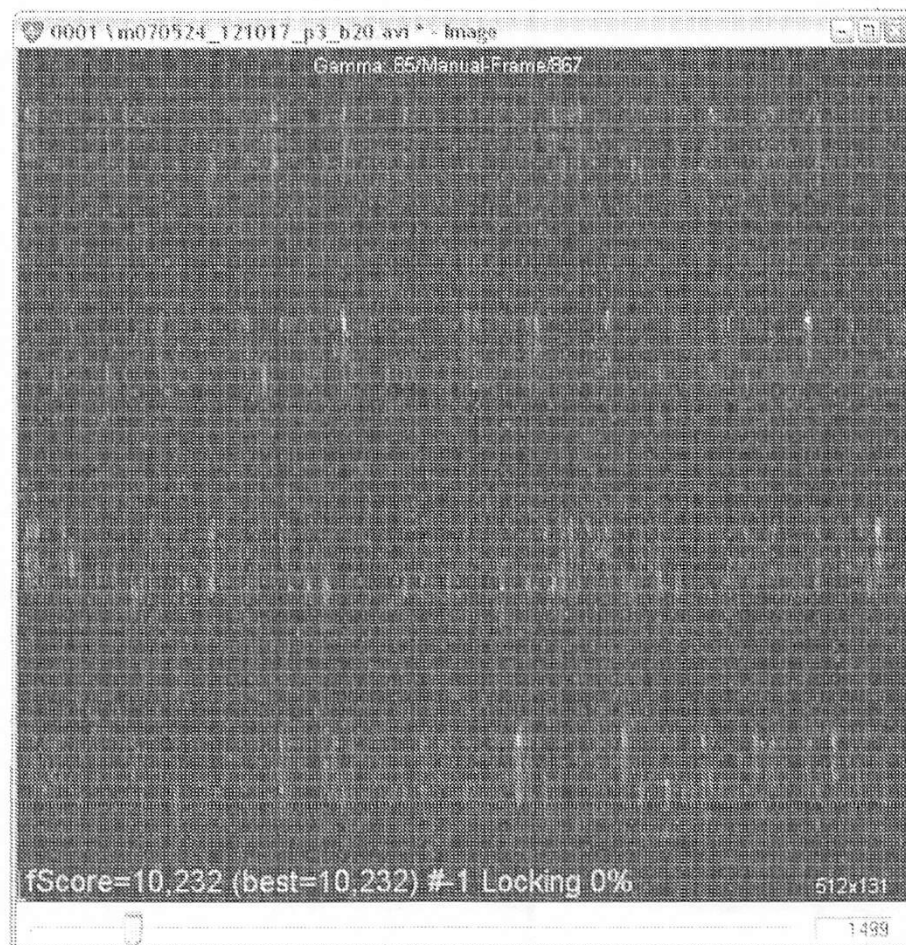
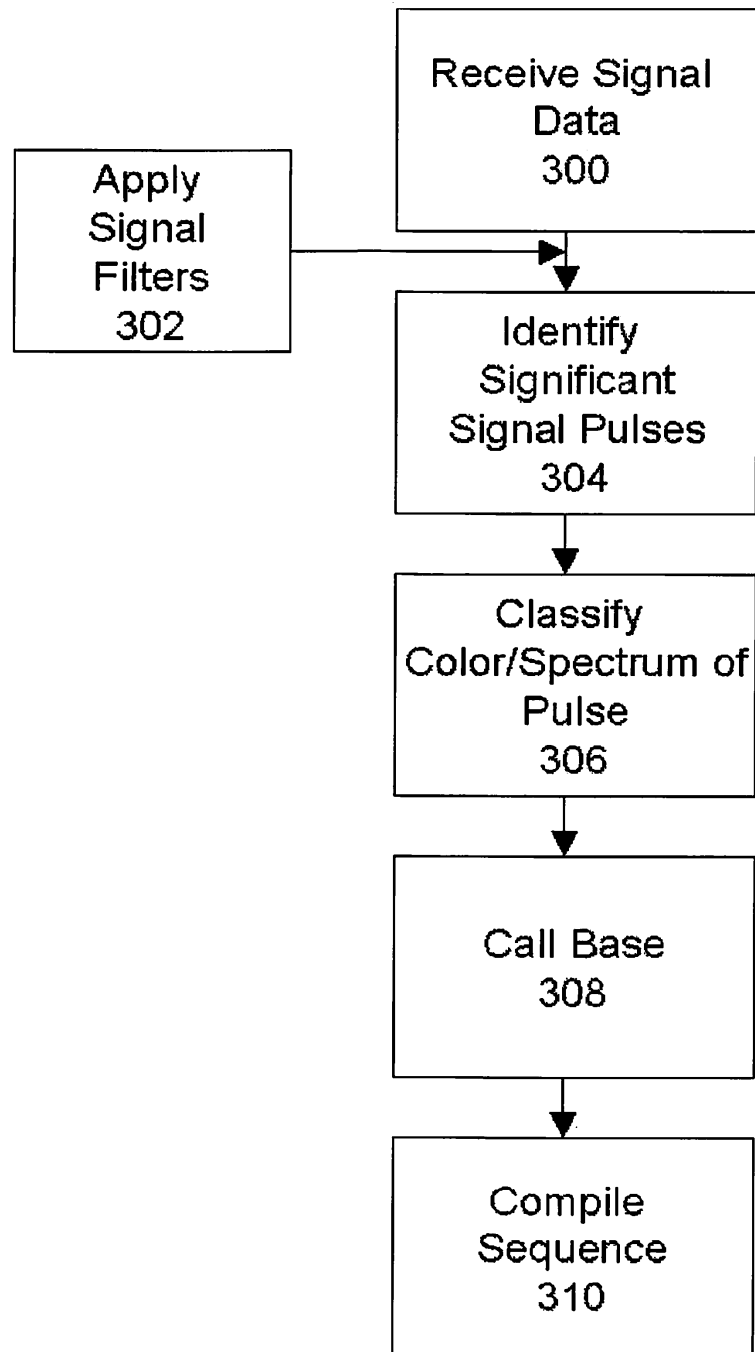
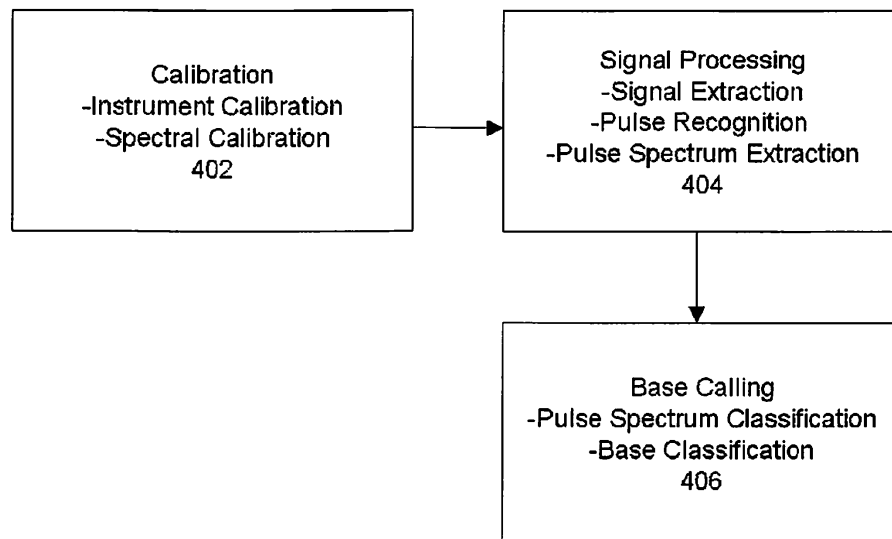
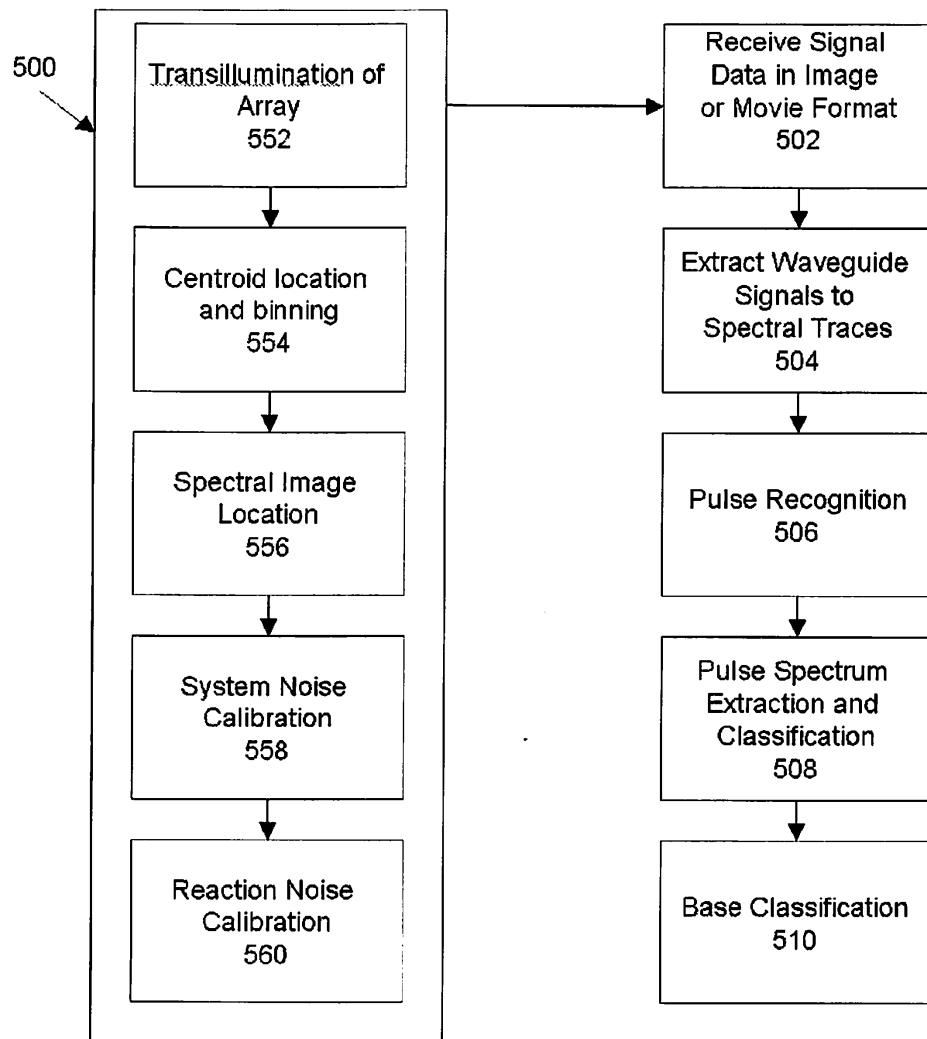


FIGURE 2

**FIGURE 3**

**FIGURE 4**

**FIGURE 5**

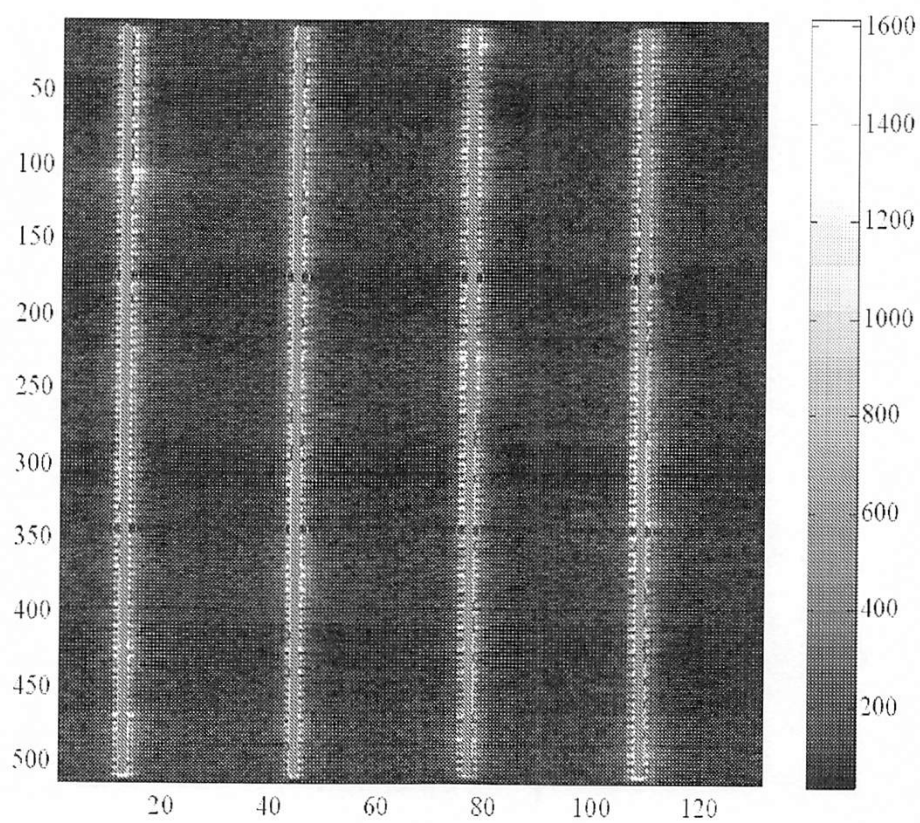


FIGURE 6

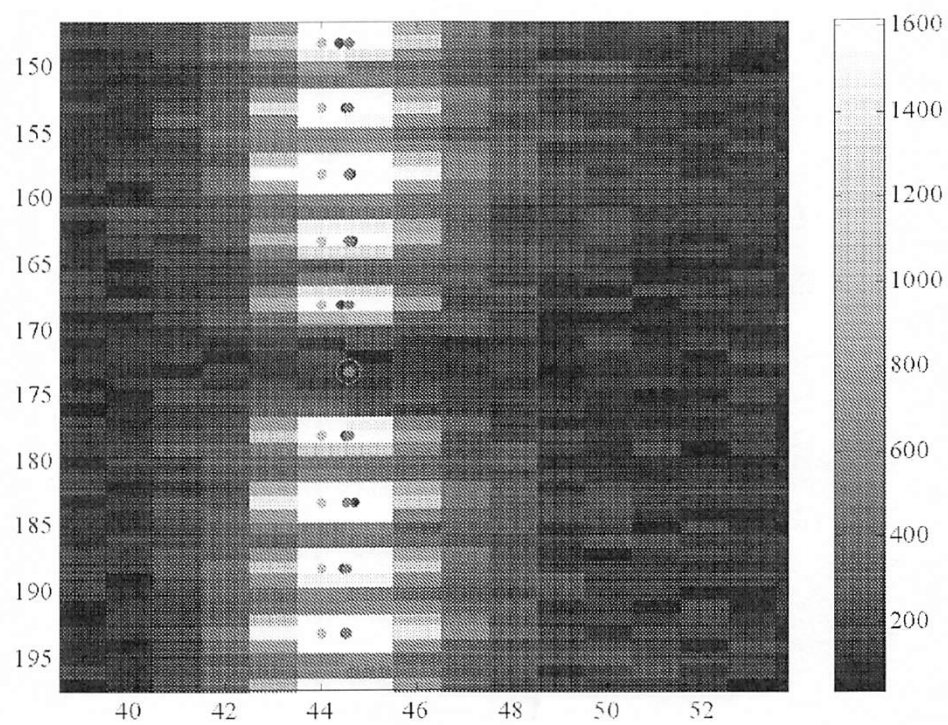


FIGURE 7

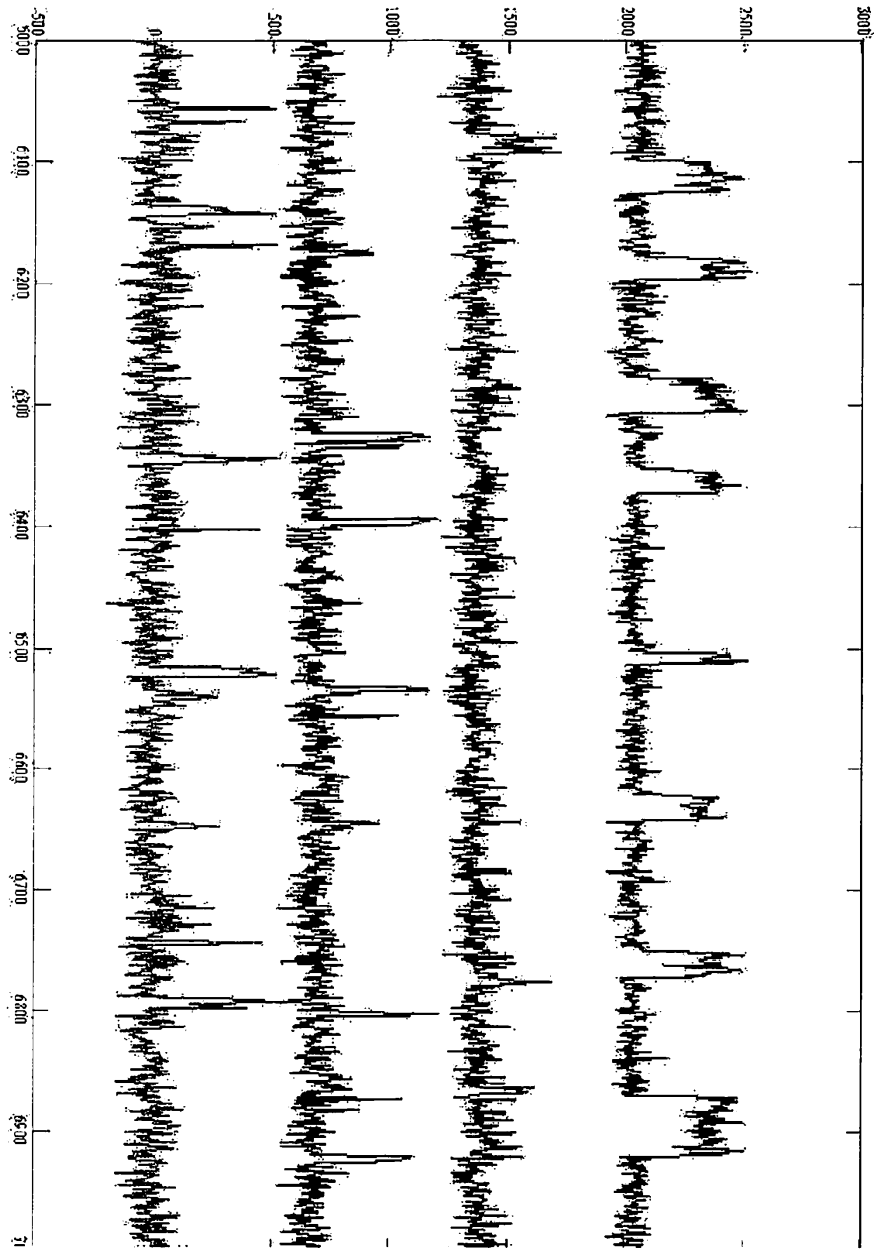
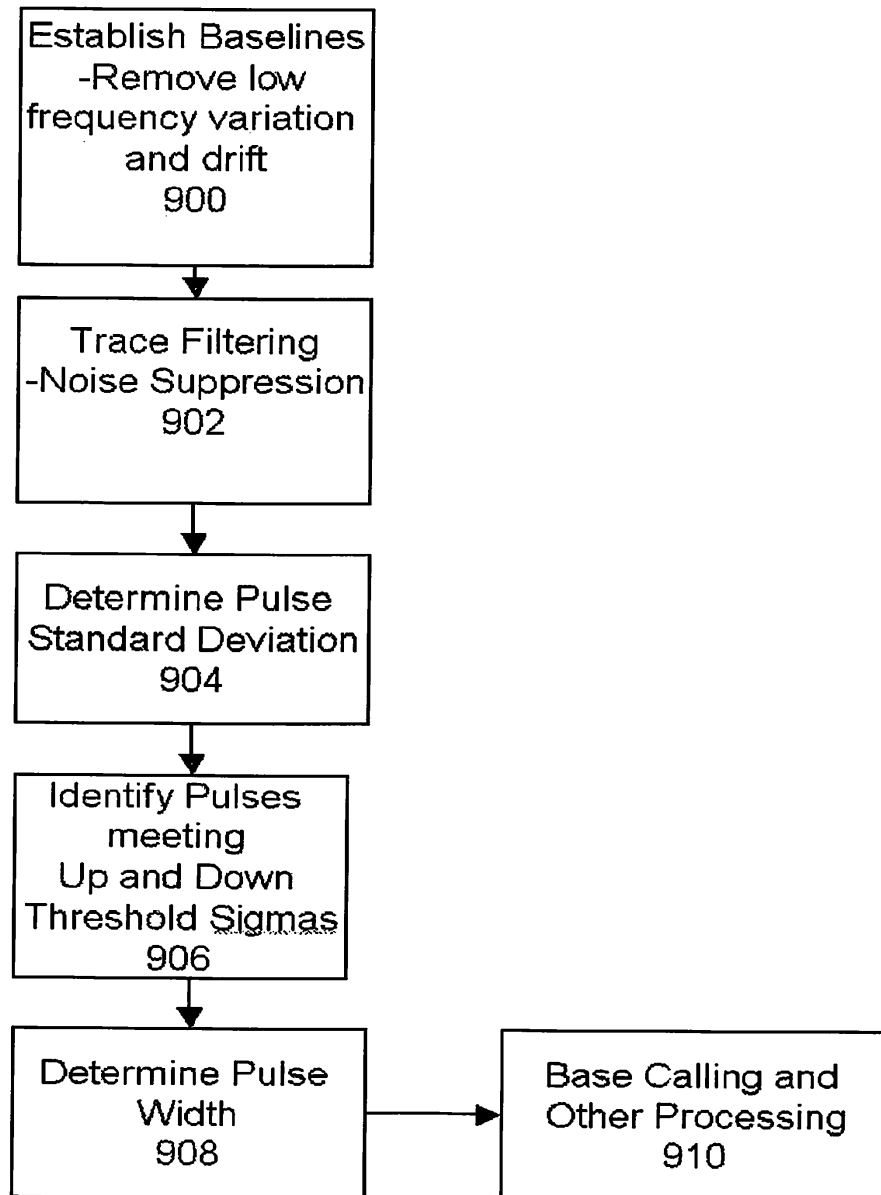
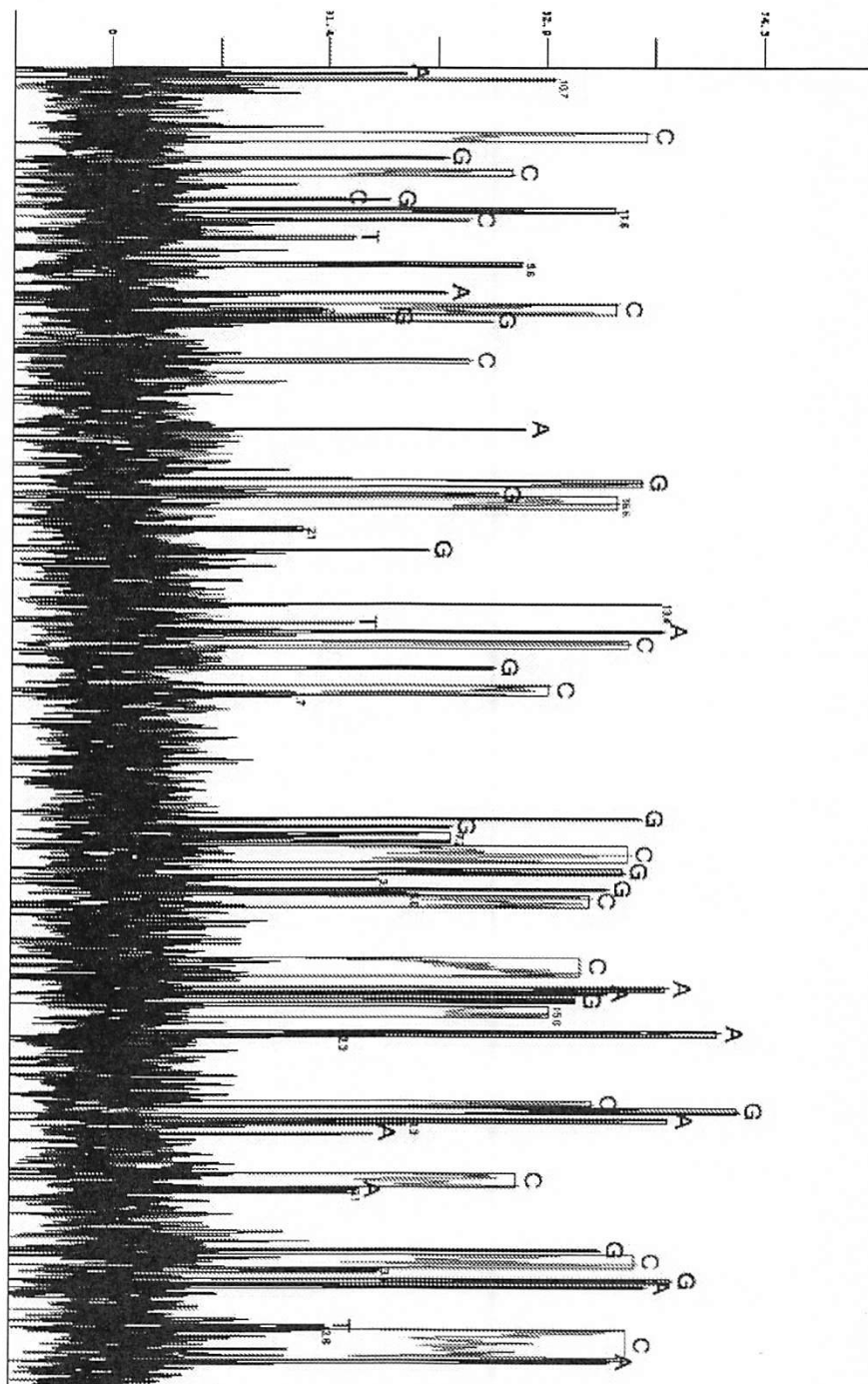


FIGURE 8

**FIGURE 9**



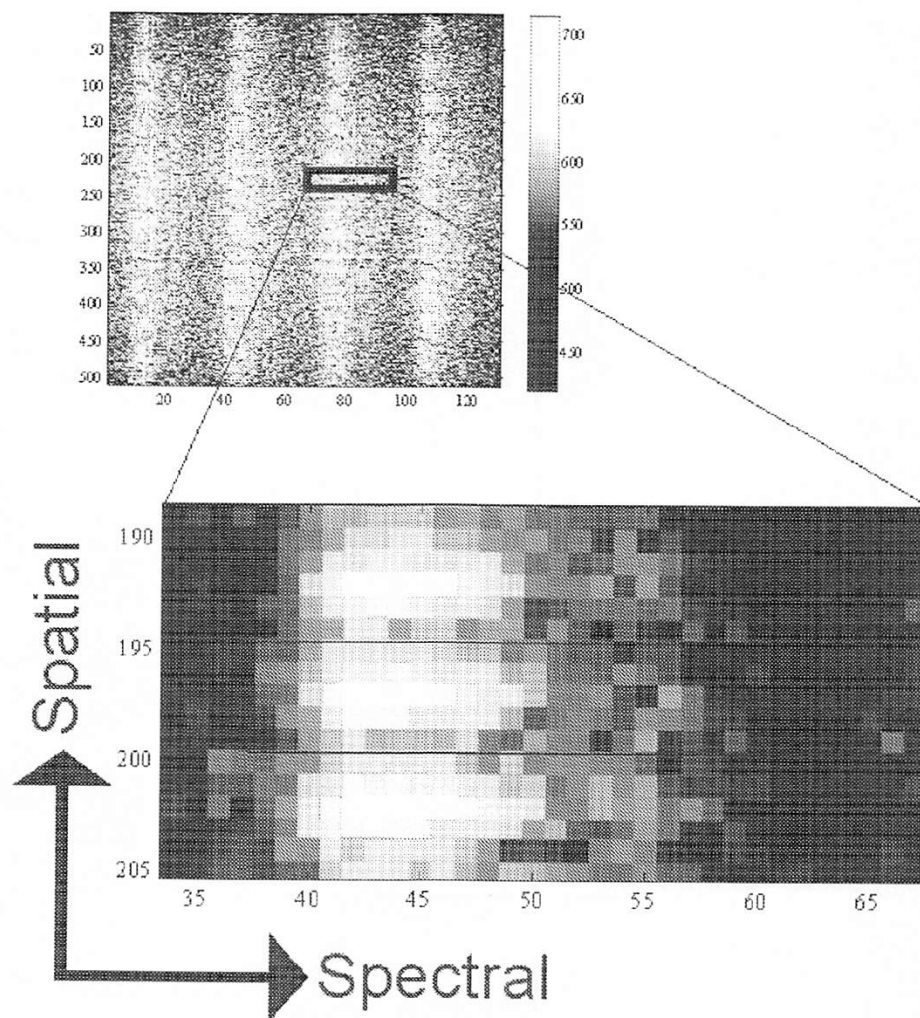


FIGURE 11

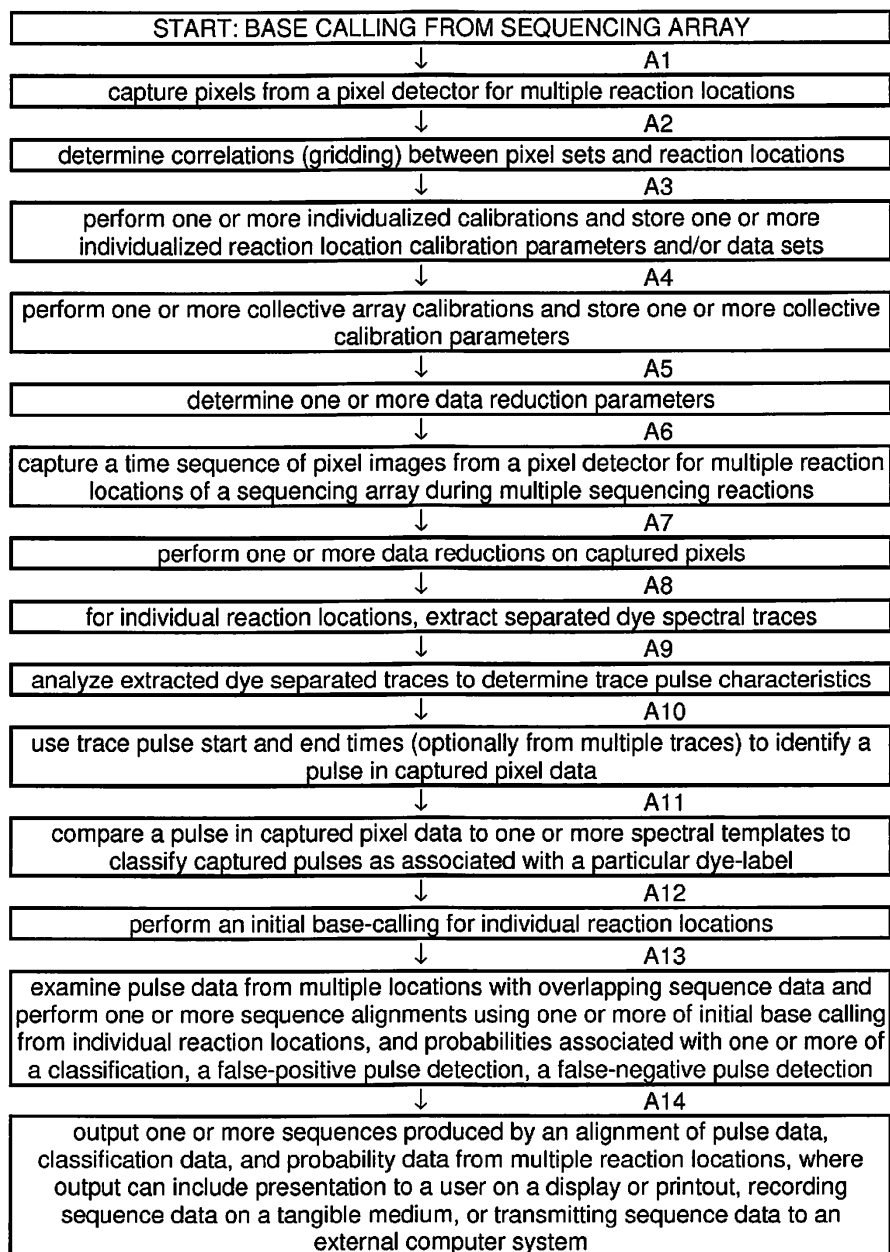


FIGURE 12

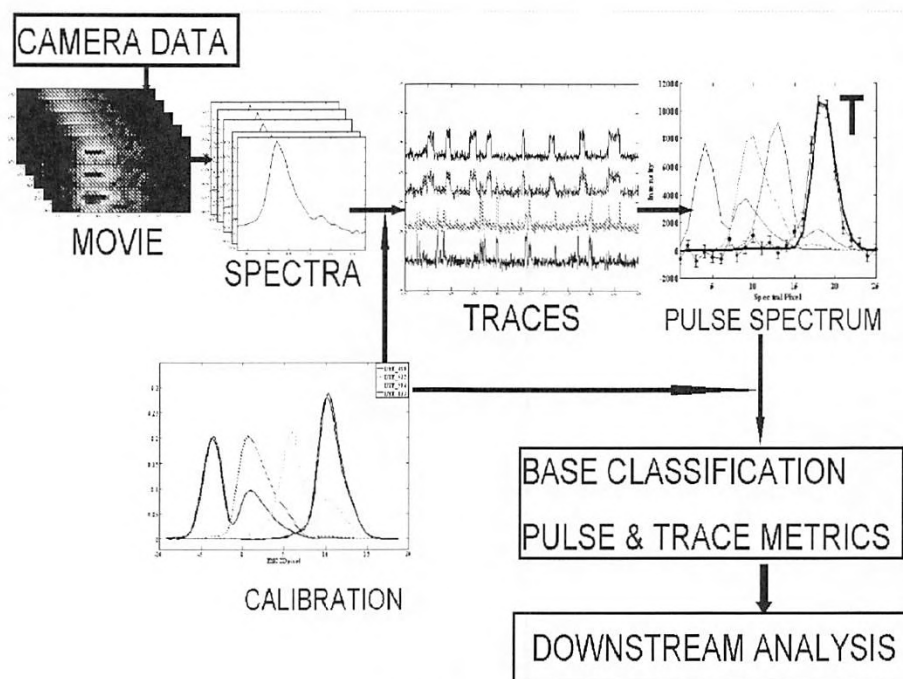
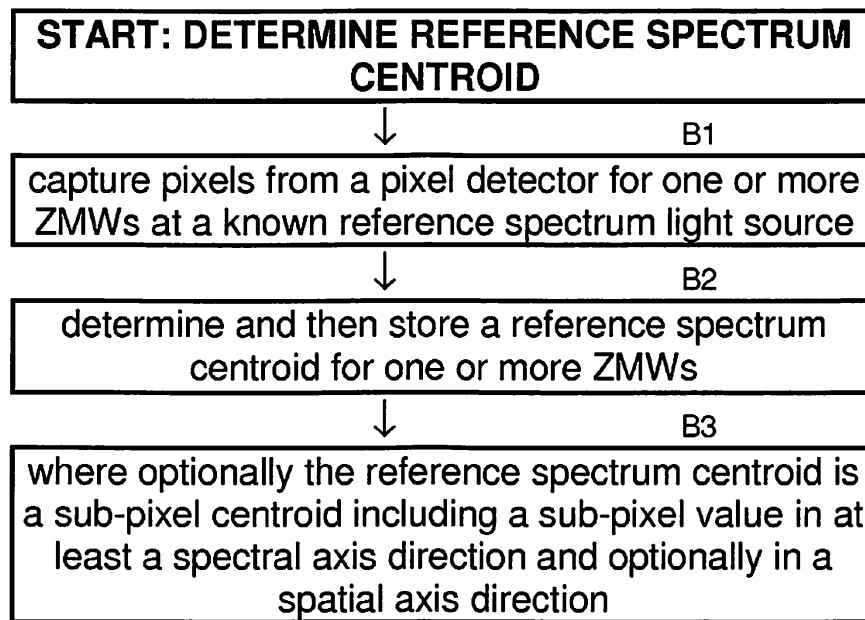


FIGURE 13

**FIGURE 14**

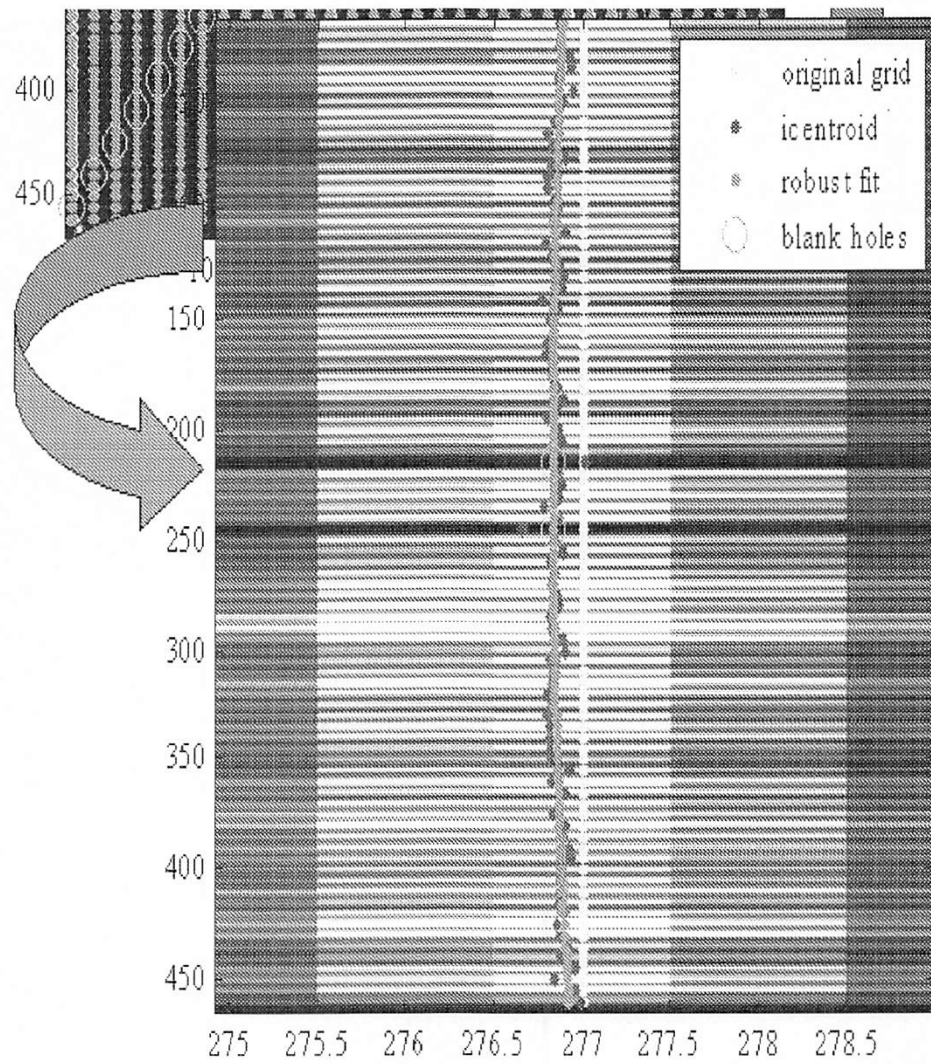


FIGURE 15

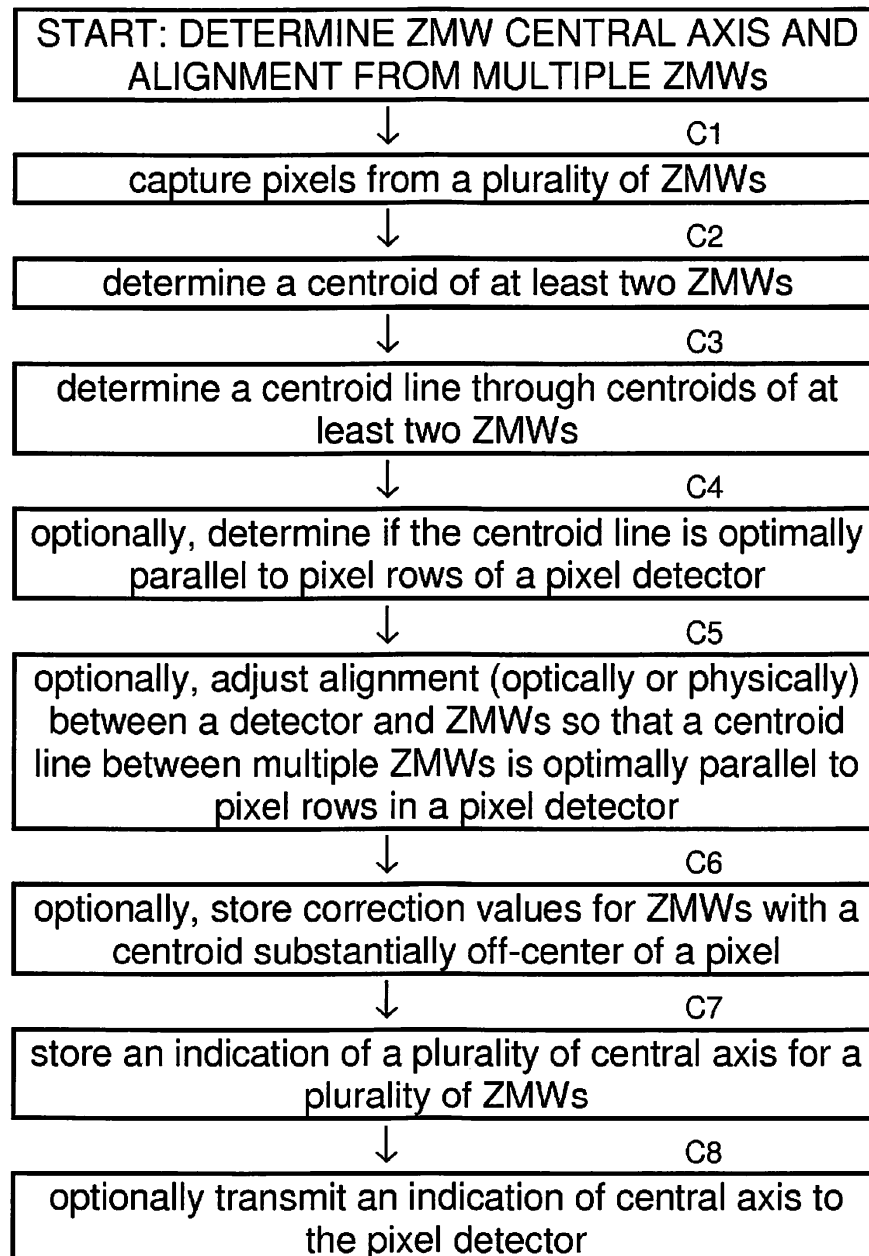
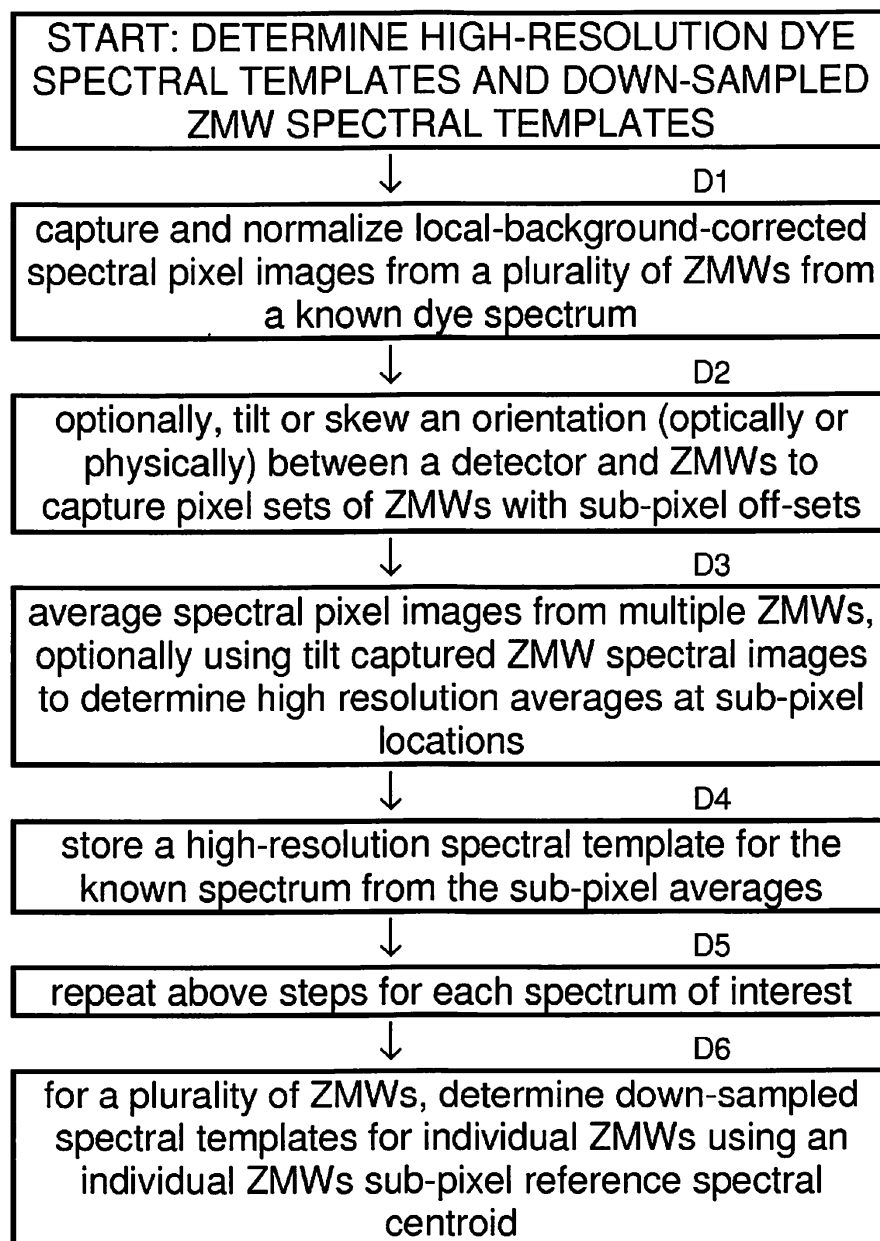


FIGURE 16

**FIGURE 17**

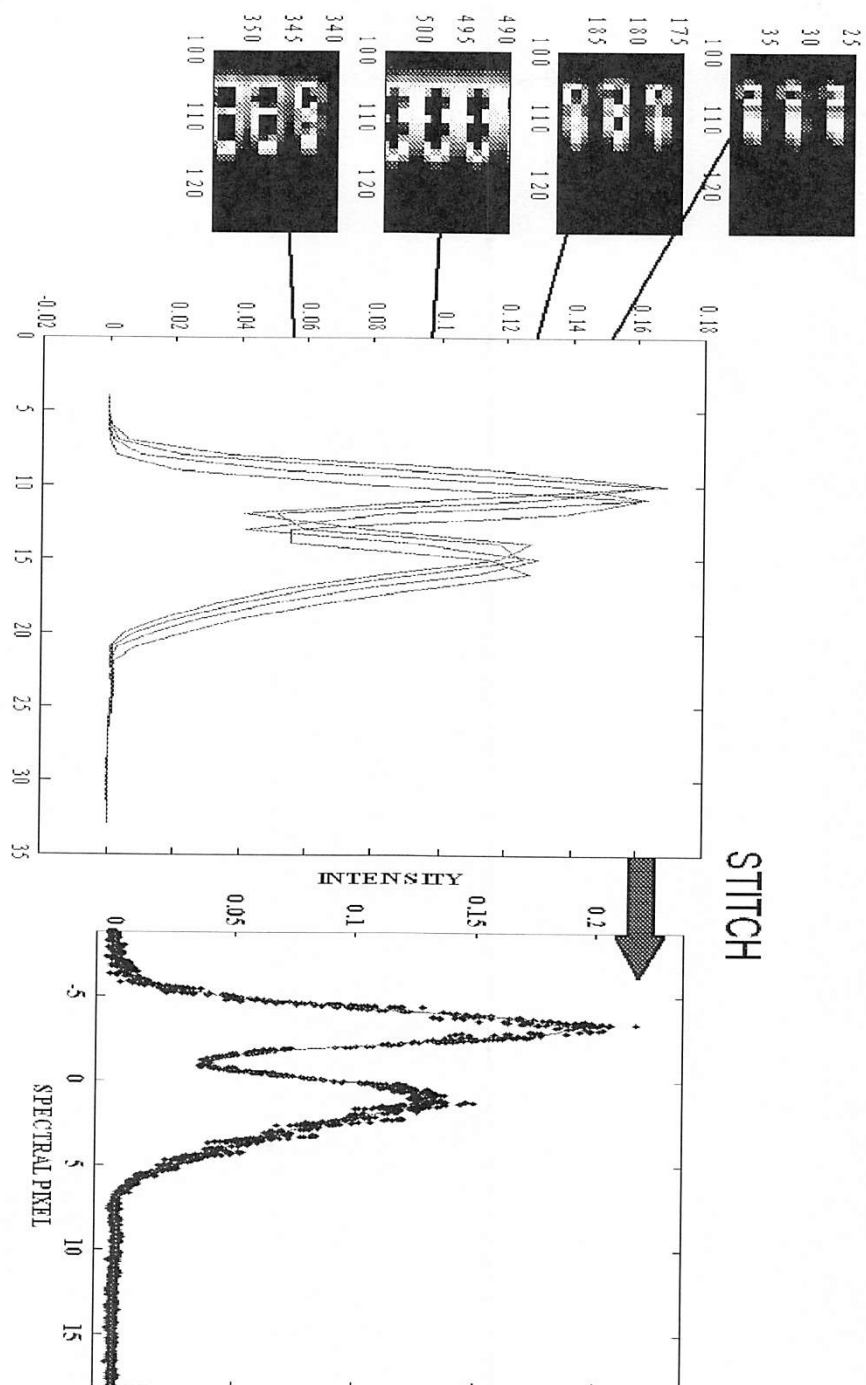


FIGURE 18

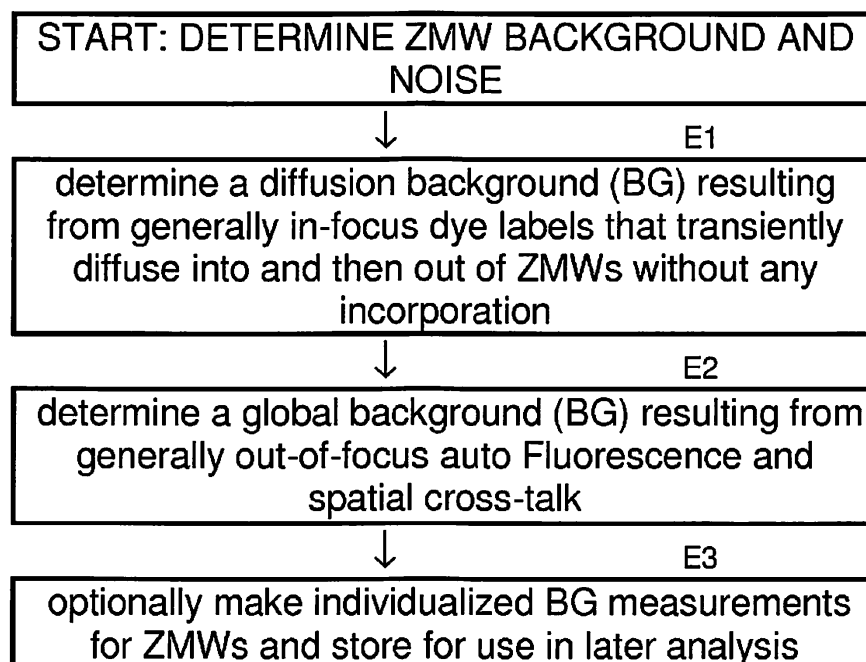


FIGURE 19

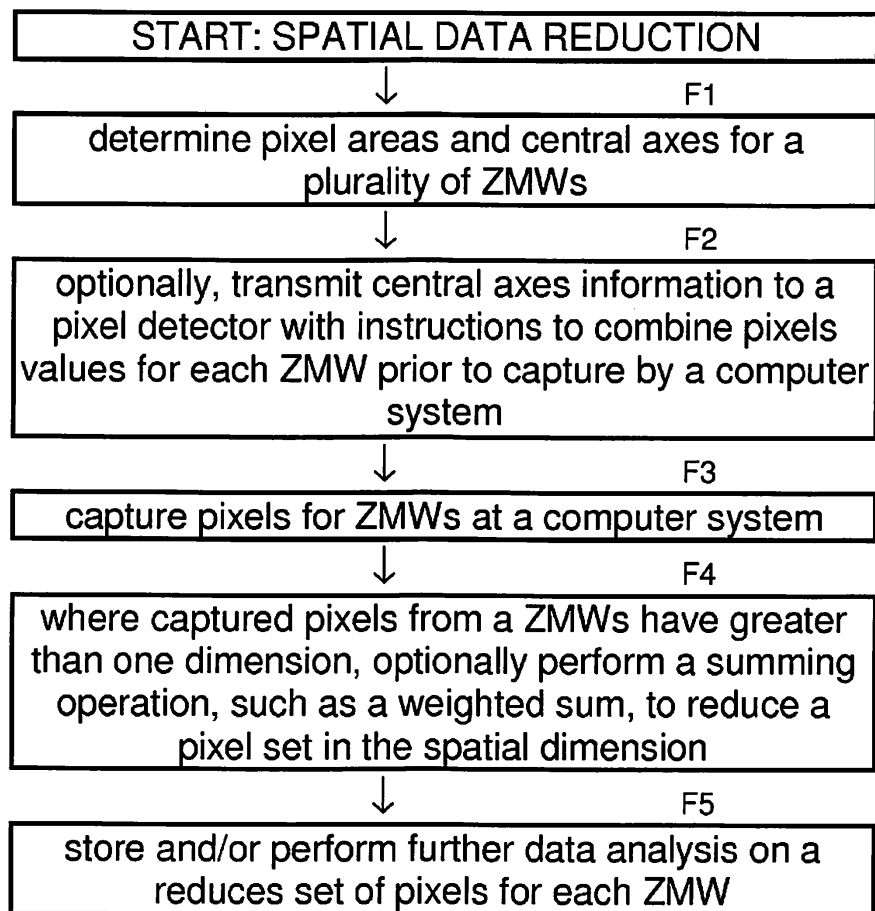


FIGURE 20

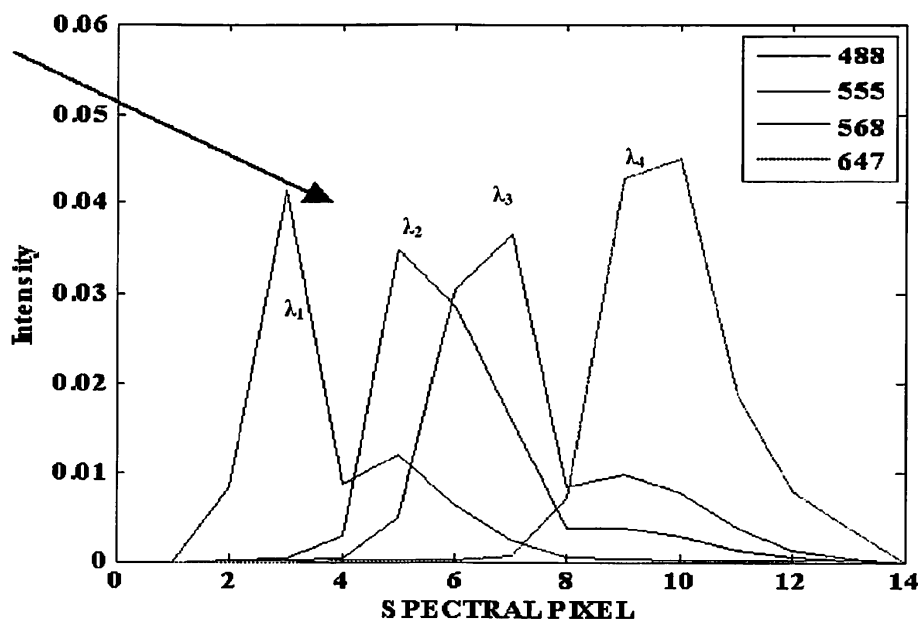
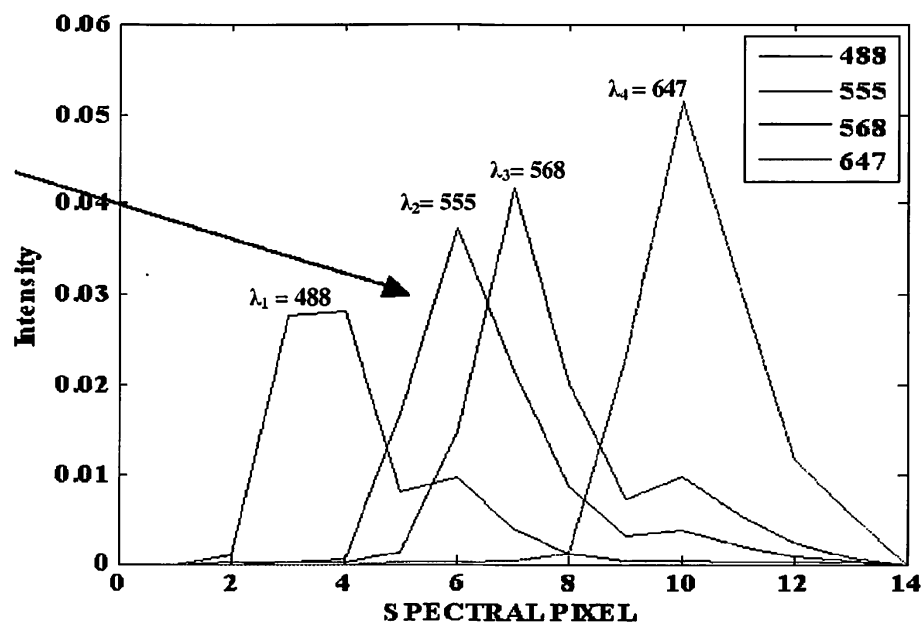
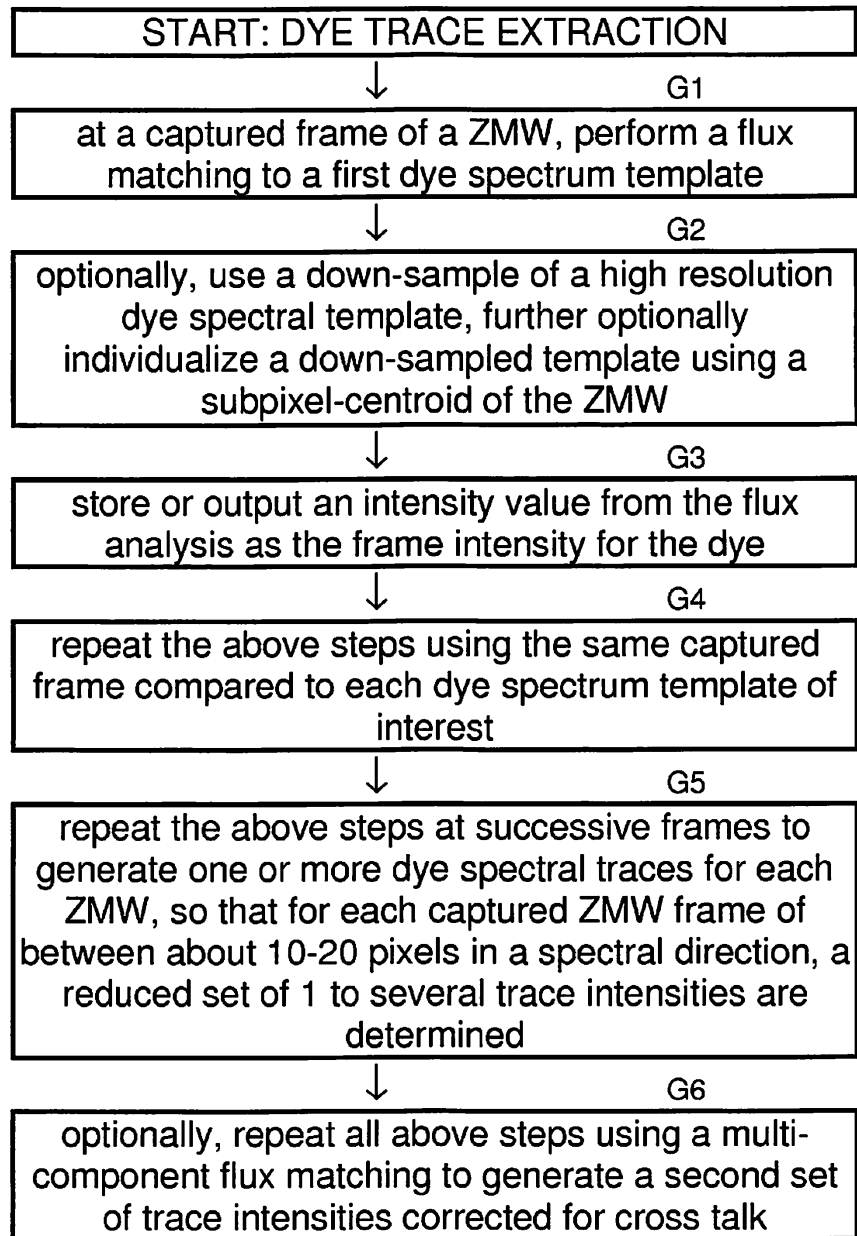


FIGURE 21

**FIGURE 22**

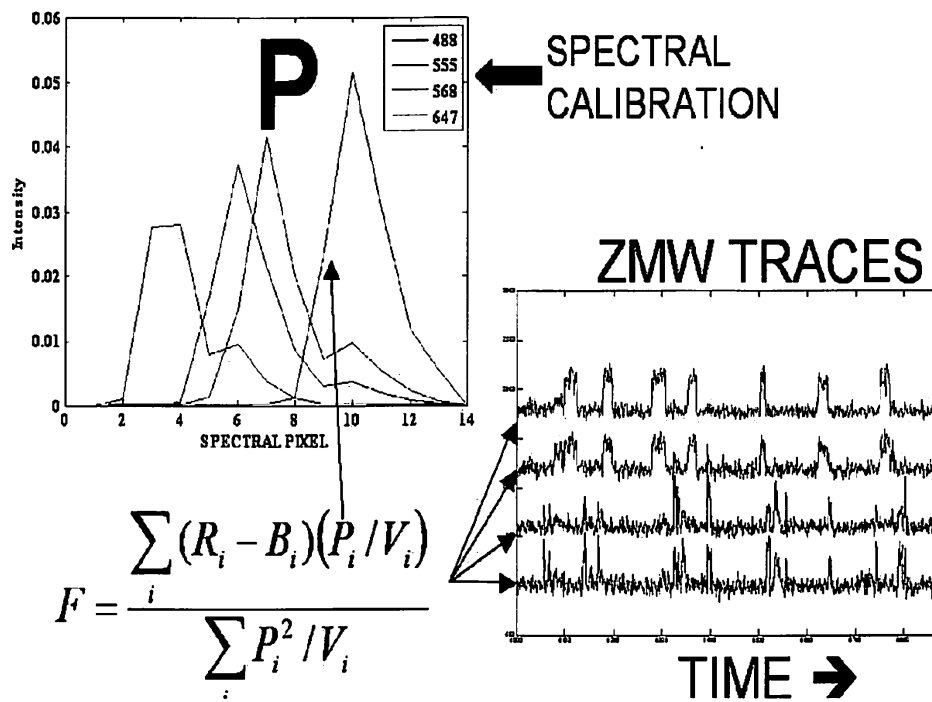


FIGURE 23

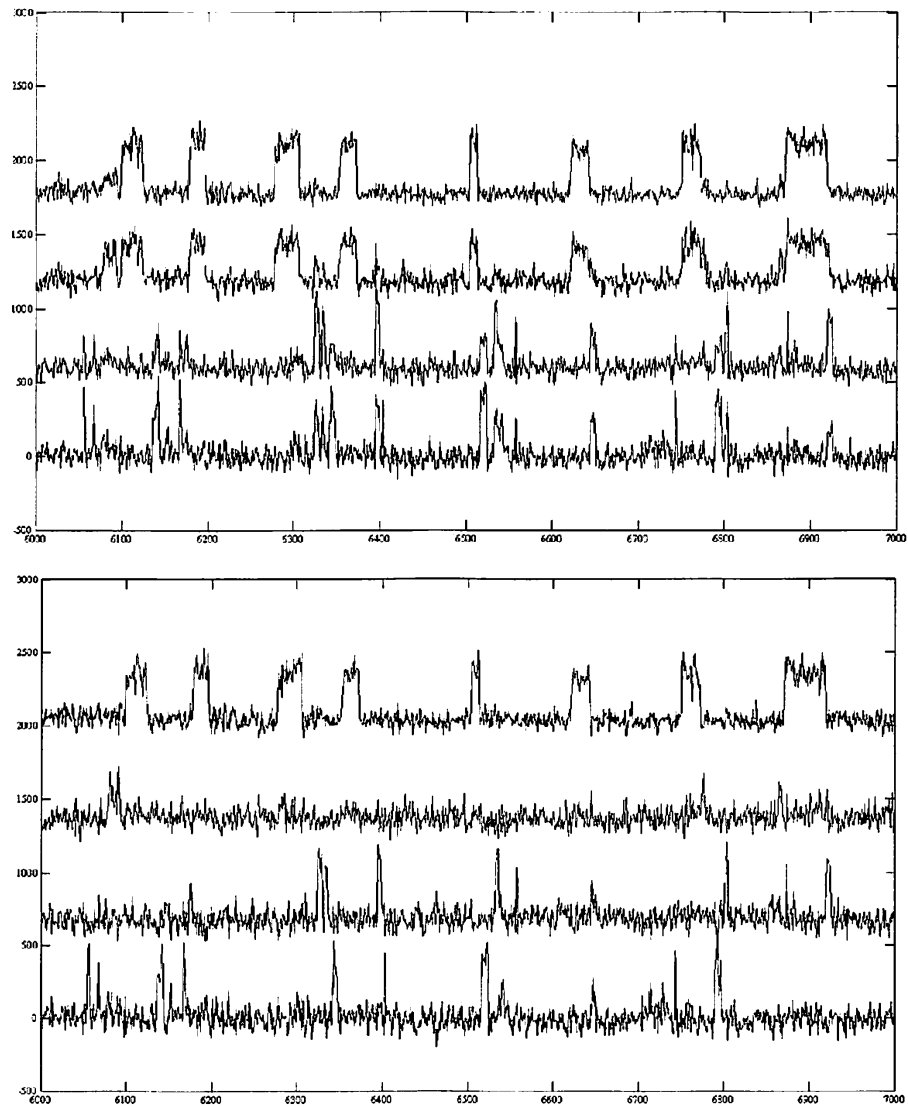


FIGURE 24

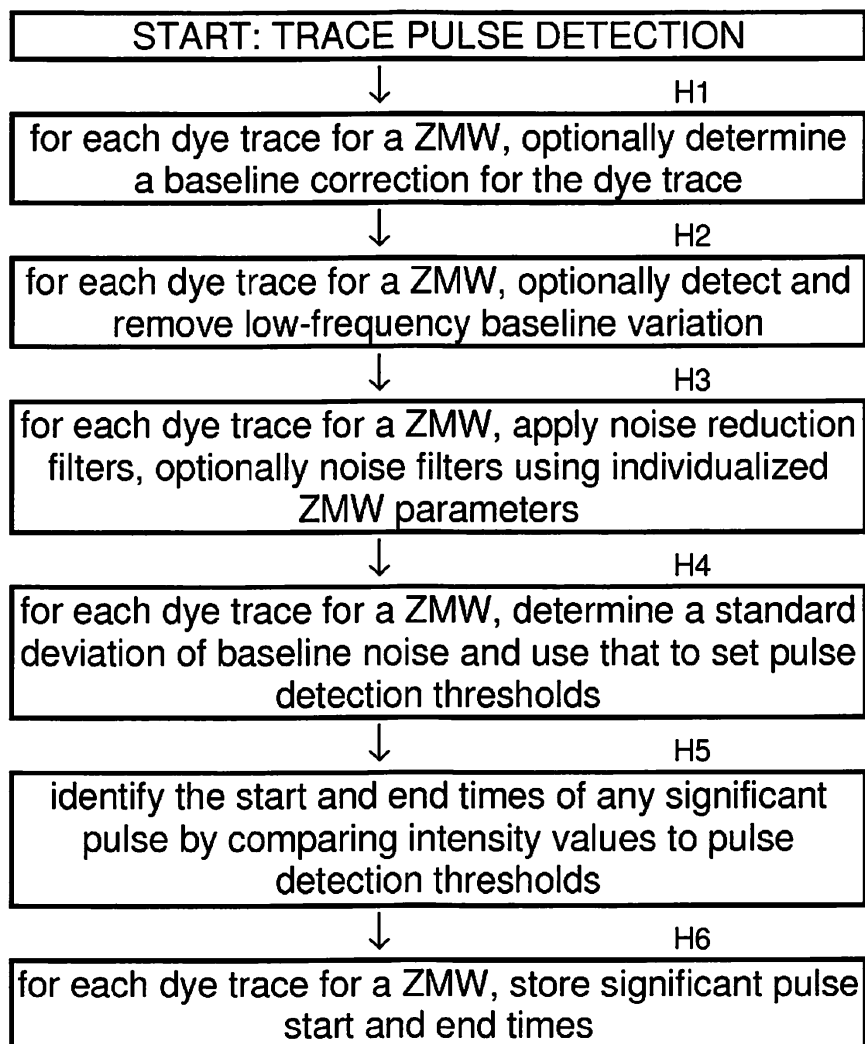


FIGURE 25

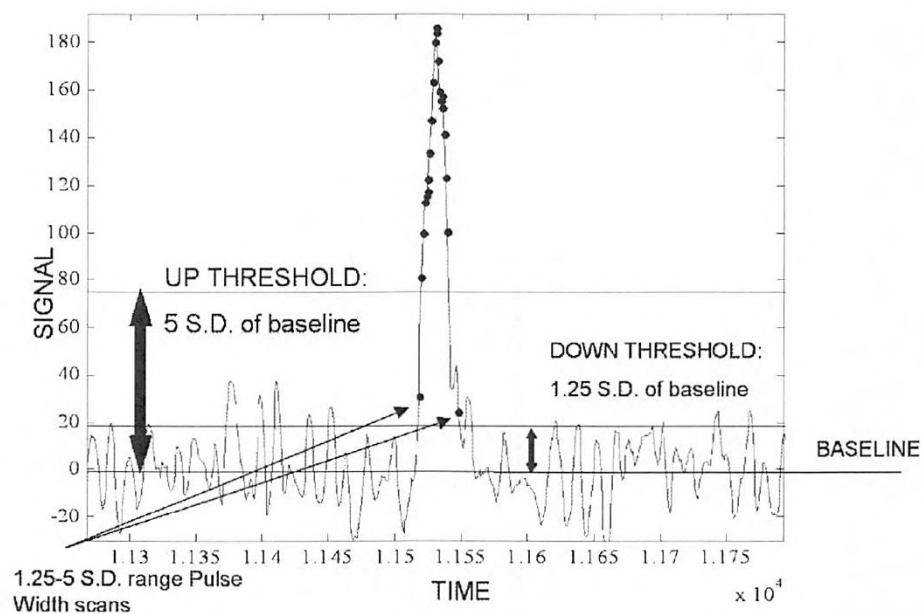


FIGURE 26

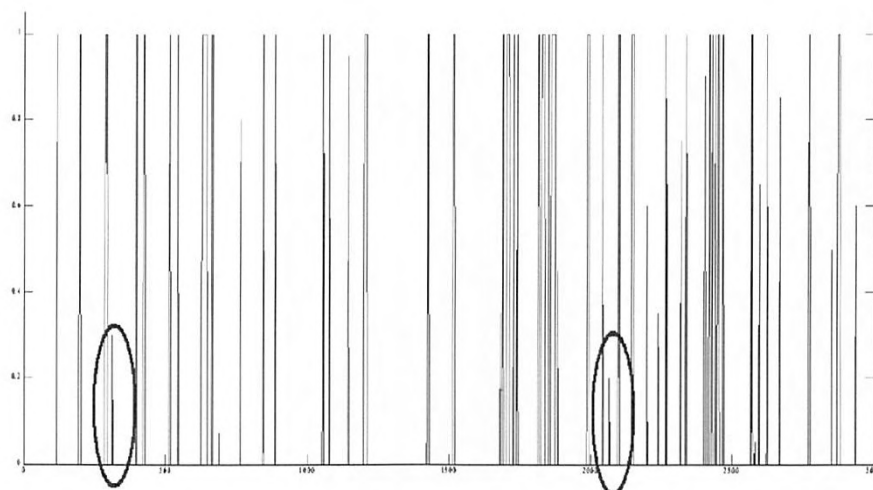


FIGURE 27

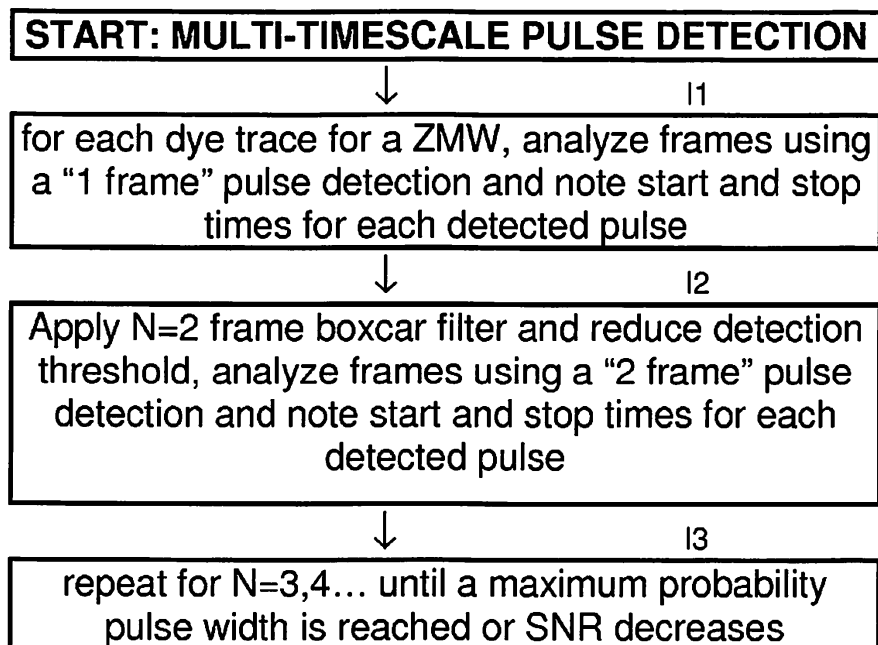


FIGURE 28

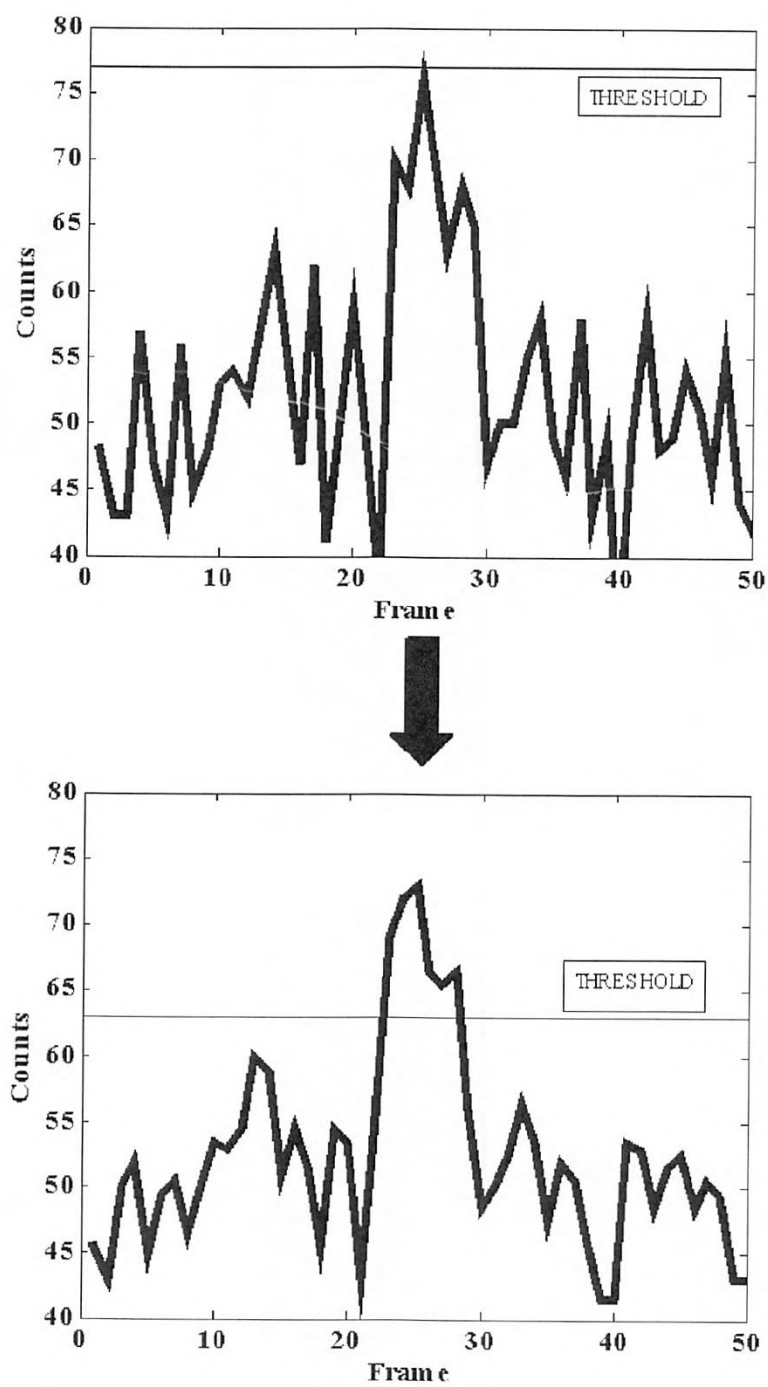


FIGURE 29

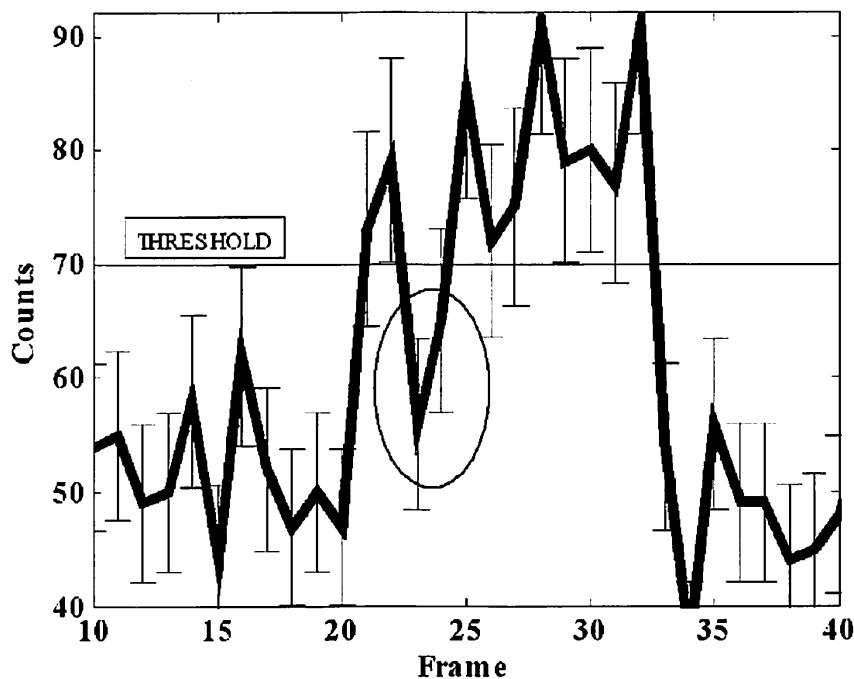


FIGURE 30

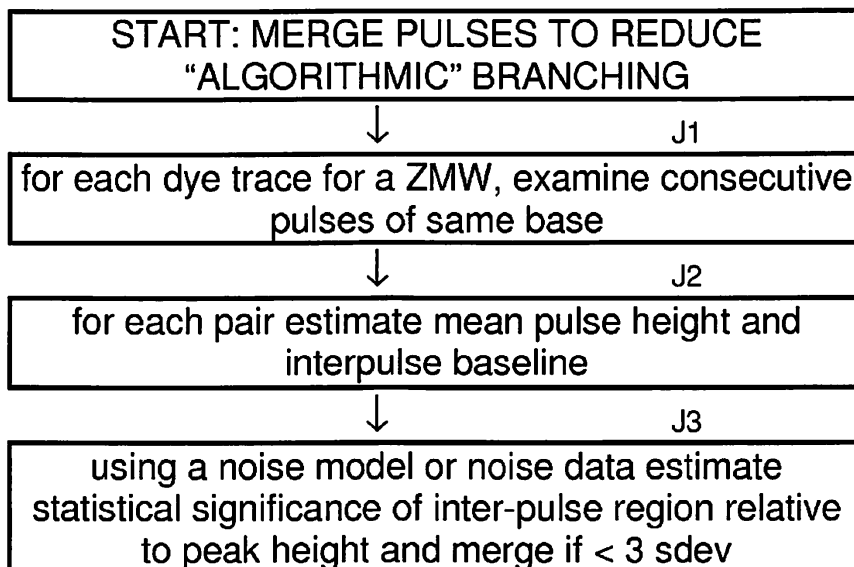


FIGURE 31

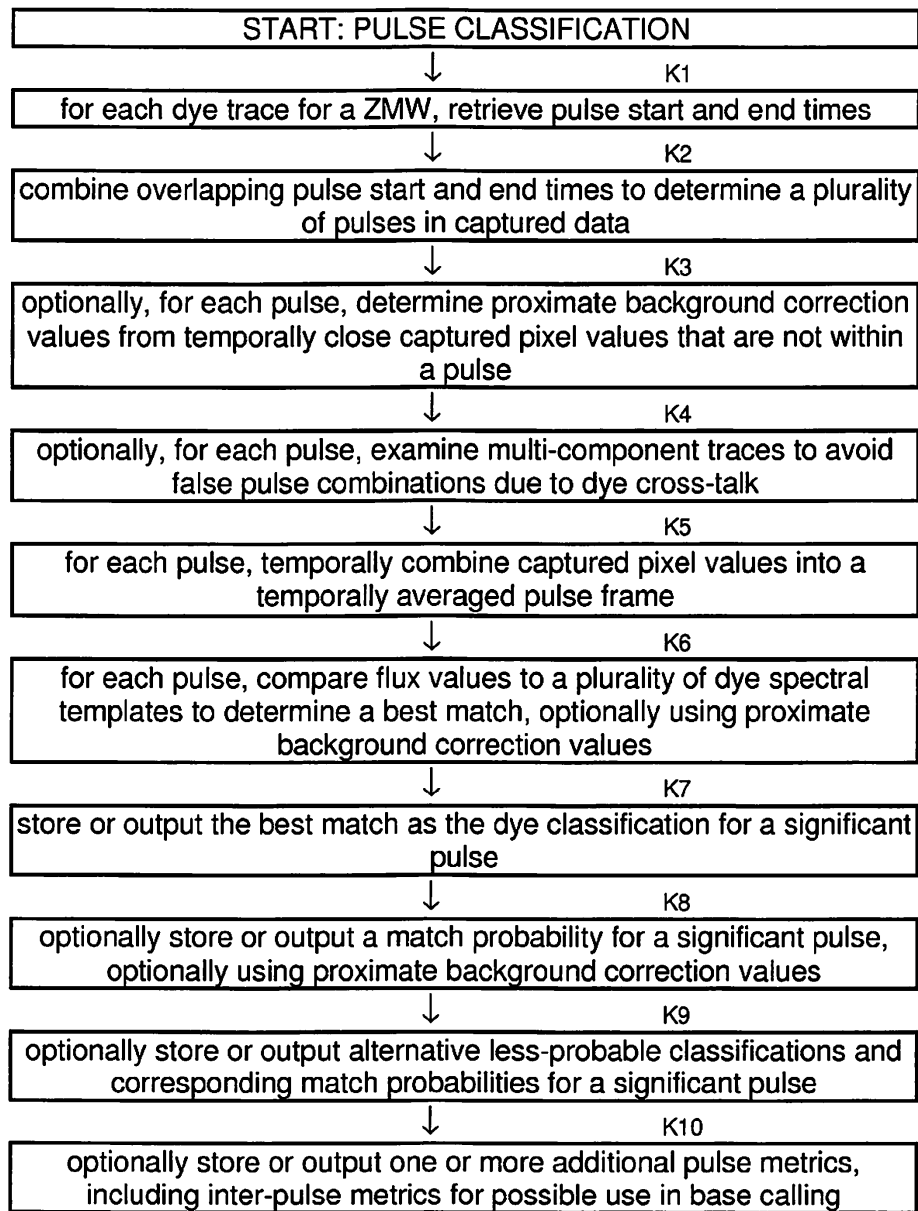


FIGURE 32

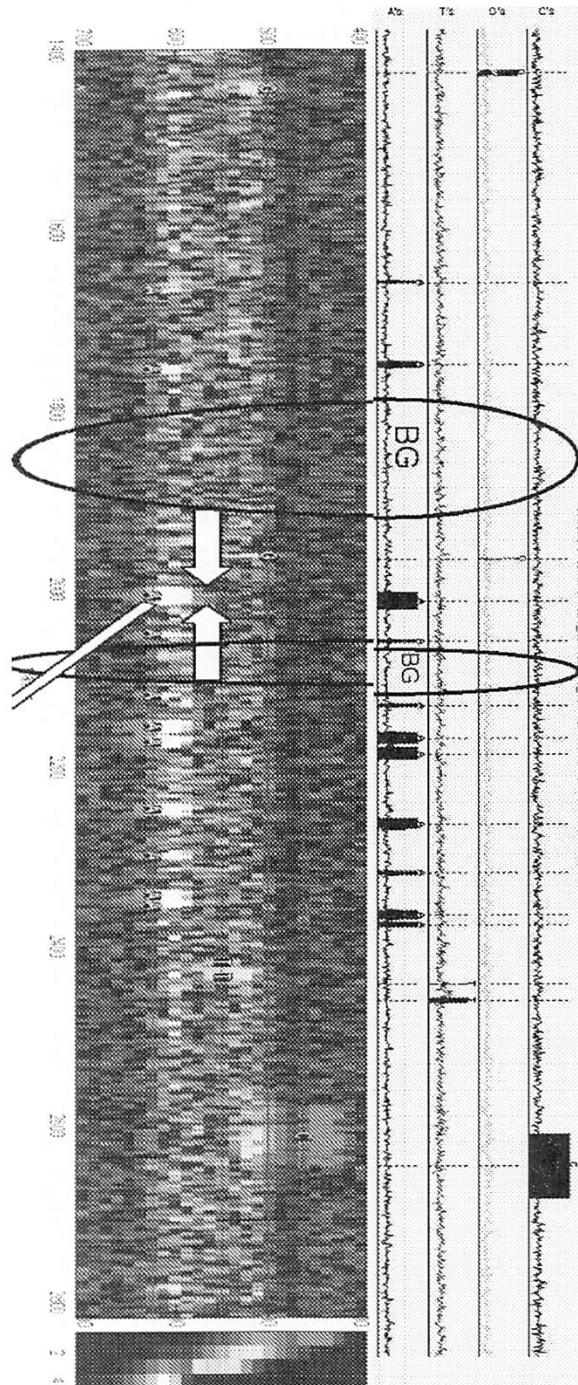


FIGURE 33A

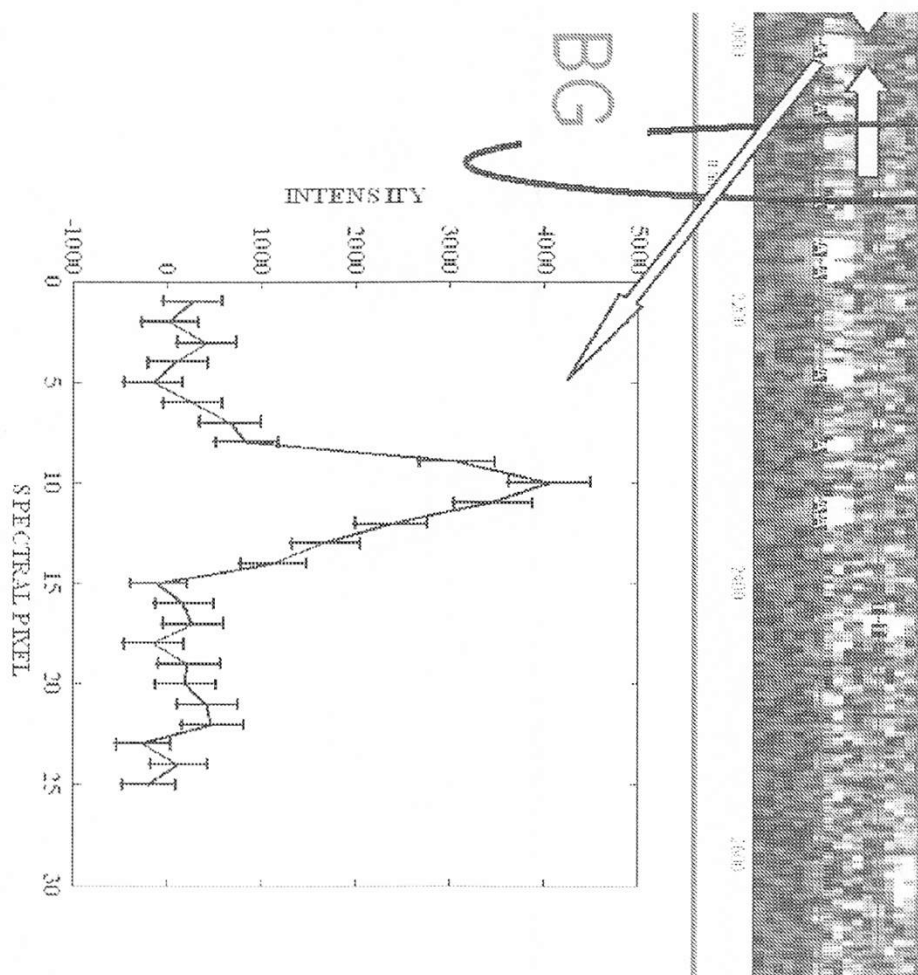
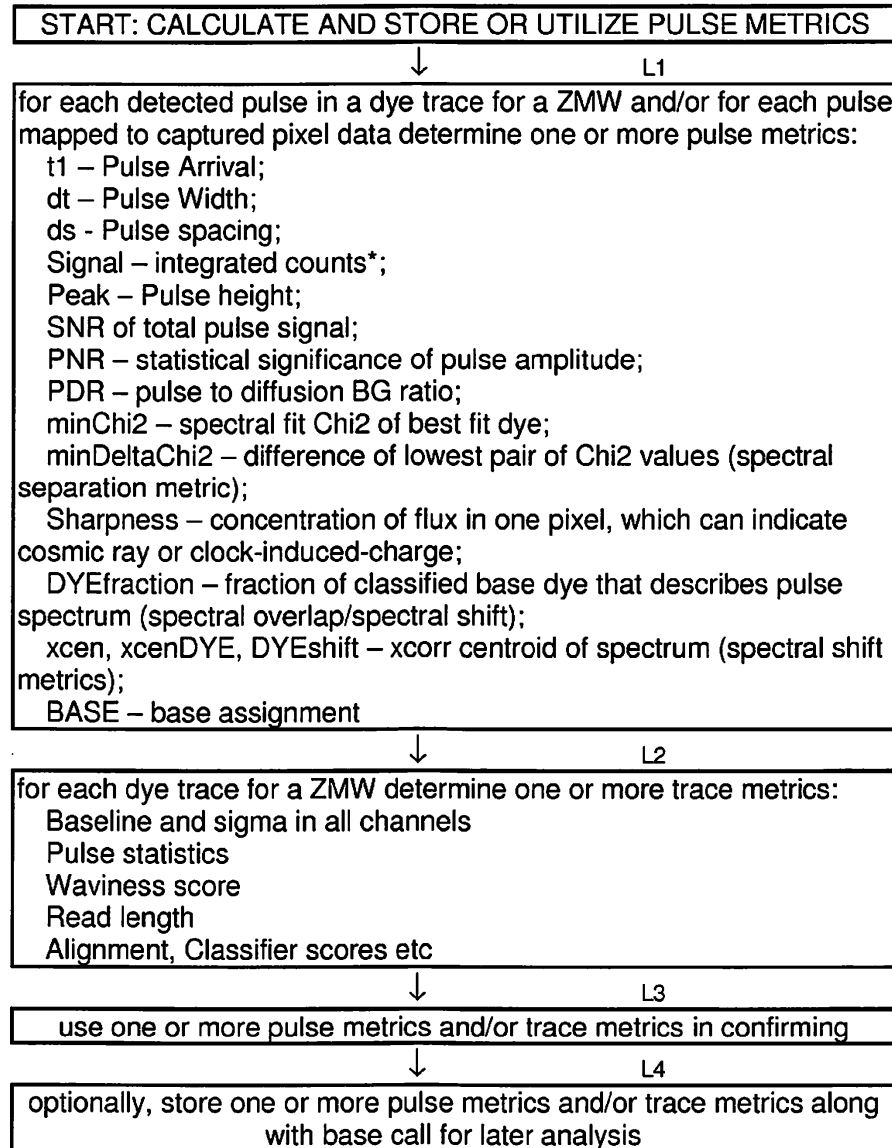


FIGURE 33B

**FIGURE 34**

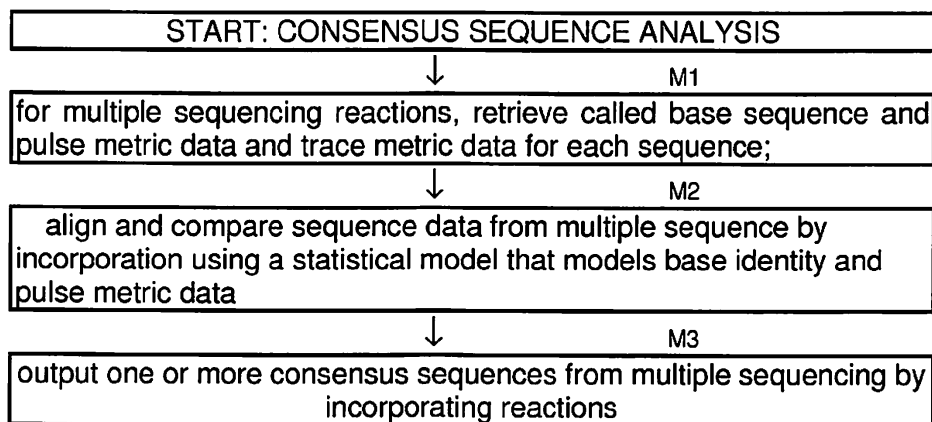


FIGURE 35

Transition Matrix: $a_{ij} = P(q_t = S_j | q_{t-1} = S_i)$

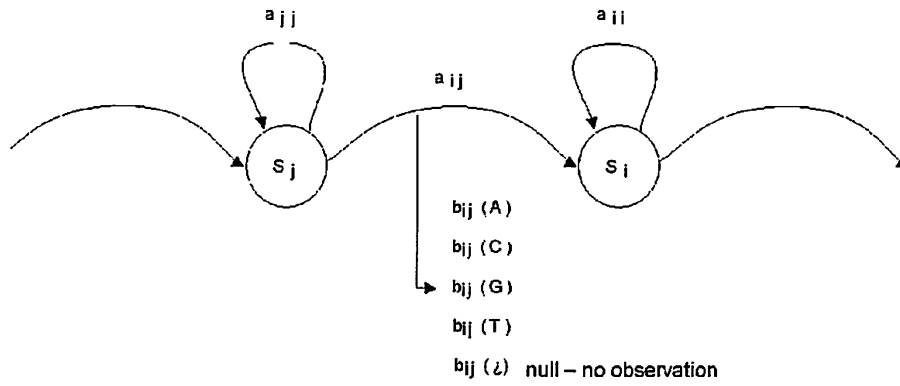


FIGURE 36

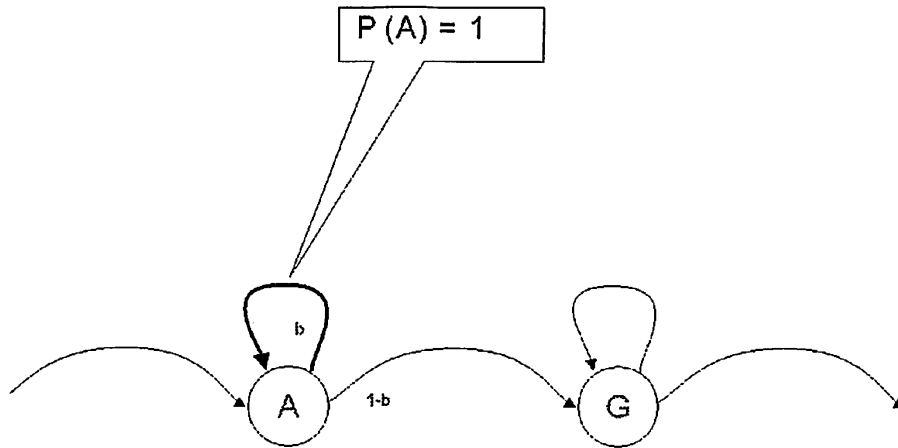


FIGURE 37

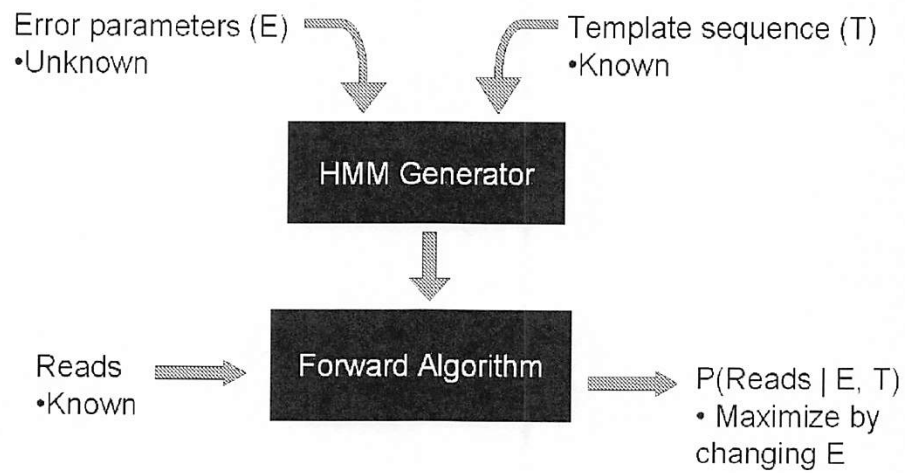


FIGURE 38

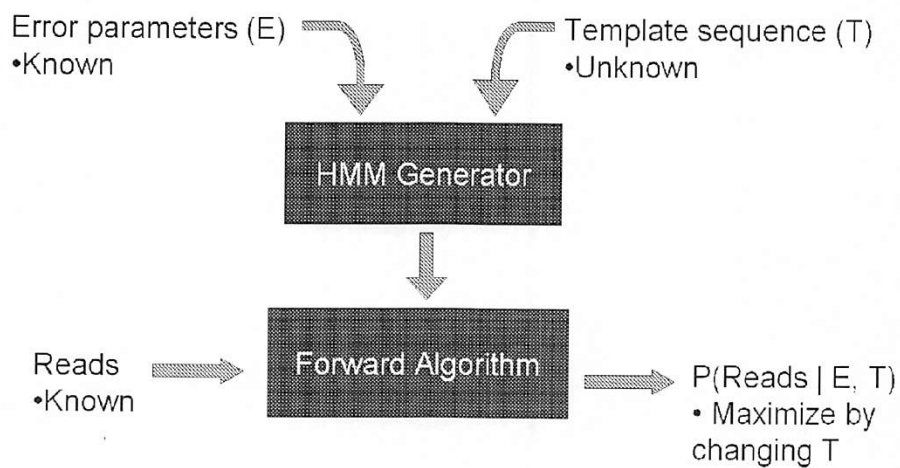


FIGURE 39

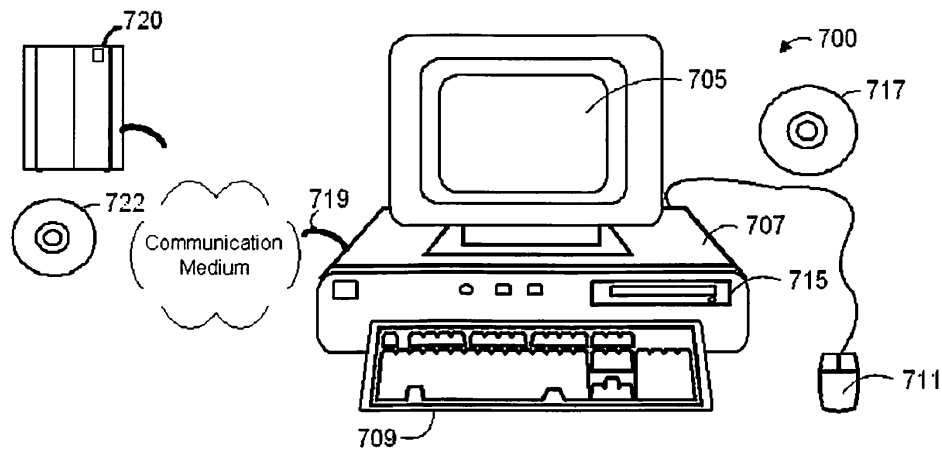


FIGURE 40

METHODS AND PROCESSES FOR CALLING BASES IN SEQUENCE BY INCORPORATION METHODS

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority from provisional patent application 60/933,399 filed 6 Jun. 2007 and incorporated herein by reference.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH

[0002] Not Applicable.

BACKGROUND OF THE INVENTION

[0003] Analysis of the subtleties of the voluminous amounts of genetic information will continue to have profound effects on the personalization of medicine. For example, this advanced genetic knowledge of patients has and will continue to have broad impact on the ability to diagnose diseases, identify predispositions to diseases or other genetically impacted disorders, the ability to identify reactivity to given drugs or other treatments, whether adverse or beneficial.

[0004] Before one can begin to interpret genetic data from patients, one must first obtain the genetic information from that patient. Technologies have been developed that allow for broad screening of large swaths of a patient's genetic code by identifying key signposts in that code and using this fragmented data as a general interpretation mechanism, e.g., using libraries of known genetic variations, such as SNPs or other polymorphisms, and correlating the profile of such variations against profiles that have a suspected association with a given disease or other phenotype.

[0005] Rather than rely upon disparate pieces of information from the genetic code, it would be of far more value to be able to rely upon the entire text of a patient's genetic code in making any interpretations from that code. In using an analogy of a novel, one gains a substantially deeper understanding of all the elements of the novel, e.g., plot, characters, setting etc., by reading the entire text, rather than by picking out individual words from disparate pages or chapters of the novel.

[0006] Technologies related to analysis of biological information have advanced rapidly over the past decade. In particular, with the improved ability to characterize genetic sequence information, identify protein structure, elucidate biological pathways, and manipulate any or all of these, has come the need for improved abilities to derive and process this information.

[0007] In the field of genetic analysis, for example, faster and faster methods of obtaining nucleic acid sequence information have consequences in terms of requiring different and often times better methods and processes for processing the raw genetic information that is generated by these processes. This progress has been evidenced in the improvements applied to separations based Sanger sequencing, where improvements in throughput and read-length have come not only through multiplexing of multi-capillary systems, but also from improvements in base calling processes that are applied to the data derived from the capillary systems.

[0008] With shifts in the underlying technology surrounding genetic analysis, also comes a necessity for a shift in the

methods and processes for processing the information from these systems. The present invention provides solutions to these and other problems.

[0009] Various embodiments and components of the present invention employ pulse, signal, and data analysis techniques that are familiar in a number of technical fields. For clarity of description, details of known techniques are not provided herein. These techniques are discussed in a number of available references works, such as: R. B. Ash, *Real Analysis and Probability*, Academic Press, New York, 1972; D. T. Bertsekas and J. N. Tsitsiklis, *Introduction to Probability*, 2002; K. L. Chung, *Markov Chains with Stationary Transition Probabilities*, 1967; W. B. Davenport and W. L. Root, *An Introduction to the Theory of Random Signals and Noise*, McGraw-Hill, New York, 1958; S. M. Kay, *Fundamentals of Statistical Processing*, Vols. 1-2, (Hardcover—1998); Monsoon H. Hayes, *Statistical Digital Signal Processing and Modeling*, 1996; *Introduction to Statistical Signal Processing* by R. M. Gray and L. D. Davisson; *Modern Spectral Estimation: Theory and Application/Book and Disk* (Prentice-Hall Signal Processing Series) by Steven M. Kay (Hardcover—January 1988); *Modern Spectral Estimation: Theory and Application* by Steven M. Kay (Paperback—March 1999); *Spectral Analysis and Filter Theory in Applied Geophysics* by Burkhard Buttkus (Hardcover—May 11, 2000); *Spectral Analysis for Physical Applications* by Donald B. Percival and Andrew T. Walden (Paperback—Jun. 25, 1993); *Astronomical Image and Data Analysis* (Astronomy and Astrophysics Library) by J.-L. Starck and F. Murtagh (Hardcover—Sep. 25, 2006); *Spectral Techniques In Proteomics* by Daniel S. Sem (Hardcover—Mar. 30, 2007); *Exploration and Analysis of DNA Microarray and Protein Array Data* (Wiley Series in Probability and Statistics) by Dhammika Amarantunga and Javier Cabrera (Hardcover—Oct. 21, 2003).

BRIEF SUMMARY OF THE INVENTION

[0010] The invention is generally directed to processes, and particularly computer implemented processes for analyzing fluorescent signals from sequence by incorporation systems, and for ultimately identifying sequence information of a target nucleic acid sequence. Consequently, the invention is also directed to systems that carry out these processes.

[0011] The invention and various specific aspects and embodiments will be better understood with reference to the following drawings and detailed descriptions. In different figures, similarly numbered items are intended to represent similar functions within the scope of the teachings provided herein. In some of the drawings and detailed descriptions below, the present invention is described in terms of the important independent embodiment of a system operating on a logic processing device, such as a computer system. This should not be taken to limit the invention, which, using the teachings provided herein, can be applied to any number of logic processors working together, whether incorporated into a camera, a detector, other optical components, or other information enabled devices or logic components incorporated into laboratory or diagnostic equipment or in functional communication therewith. For purposes of clarity, this discussion refers to devices, methods, and concepts in terms of specific examples. However, the invention and aspects thereof may have applications to a variety of types of devices and systems. It is therefore intended that the invention not be limited except as provided in the attached claims.

[0012] Furthermore, it is well known in the art that logic systems and methods such as described herein can include a variety of different components and different functions in a modular fashion. Different embodiments of the invention can include different mixtures of elements and functions and may group various functions as parts of various elements. For purposes of clarity, the invention is described in terms of systems that include many different innovative components and innovative combinations of innovative components and known components. No inference should be taken to limit the invention to combinations containing all of the innovative components listed in any illustrative embodiment in this specification. The functional aspects of the invention that are implemented on a computer or other logic processing systems or circuits, as will be understood from the teachings herein, may be implemented or accomplished using any appropriate implementation environment or programming language, such as C, C++, Cobol, Pascal, Java, Java-script, HTML, XML, dHTML, assembly or machine code programming, RTL, etc. All references, publications, patents, and patent applications cited herein are hereby incorporated by reference in their entirety for all purposes.

BRIEF DESCRIPTION OF THE DRAWINGS

[0013] FIG. 1 provides a schematic illustration of an overall system used, inter alia, for sequencing by incorporation analyses.

[0014] FIG. 2 is an EMCCD derived image of fluorescent signal pulses from an array of sequencing reactions.

[0015] FIG. 3 provides a schematic flow-chart of the process used in monitoring sequencing reactions and identifying bases in the sequence.

[0016] FIG. 4 is a high level flow chart breakdown of an overall real time sequence by incorporation process.

[0017] FIG. 5 is a detailed flow chart illustrating a number of the process steps included in the overall process shown in FIG. 4.

[0018] FIG. 6 is an EMCCD image of a trans-illuminated array of zero mode waveguides.

[0019] FIG. 7 is a close up image of a subset of the trans-illuminated waveguides in the array which also shows the centroid determination for each imaged waveguide.

[0020] FIG. 8 illustrates an exemplary 4 channel signal trace derived from an imaged waveguide.

[0021] FIG. 9 provides a schematic flow-chart of a pulse recognition process.

[0022] FIG. 10 illustrates an integrated trace following base classification of a pulse trace, including base calls.

[0023] FIG. 11 is a spectral reference pixel image of a trans-illuminated array of zero mode waveguides according to specific embodiments of the invention.

[0024] FIG. 12 is a flow chart illustrating an example method of base calling from a sequencing array using a logic processing system according to specific embodiments of the invention.

[0025] FIG. 13 is a diagram illustrating an example method of base calling from a sequencing array using a logic processing system according to specific embodiments of the invention.

[0026] FIG. 14 is a flow chart illustrating an example method of determining a reference spectral centroid for a reaction location image according to specific embodiments of the invention.

[0027] FIG. 15 is a diagram illustrating an initial gridding step and corrected centroids according to specific embodiments of the invention.

[0028] FIG. 16 is a flow chart illustrating an example method of determining an alignment for reaction locations and a central axis for individual reaction locations from multiple reaction location images according to specific embodiments of the invention.

[0029] FIG. 17 is a flow chart illustrating an example method of determining high-resolution dye spectral templates and down-sampled individualized spectral templates according to specific embodiments of the invention.

[0030] FIG. 18 is a diagram illustrating stitching high-resolution dye spectral templates from multiple captured spectral calibration data according to specific embodiments of the invention.

[0031] FIG. 19 is a flow chart illustrating an example method of determining background noise for reaction location images according to specific embodiments of the invention.

[0032] FIG. 20 is a flow chart illustrating an example method of data reduction for a reaction location image according to specific embodiments of the invention.

[0033] FIG. 21 illustrates an example of four superimposed spectral images representing different dyes imaged from two different reaction locations according to specific embodiments of the invention.

[0034] FIG. 22 is a flow chart illustrating an example method of dye trace extraction from a time series of a reaction location images according to specific embodiments of the invention.

[0035] FIG. 23 is a diagram illustrating using an individualized set of spectral templates for a ZMW to extract four spectral traces using a flux calculation according to specific embodiments of the invention.

[0036] FIG. 24 is a diagram illustrating a set of dye-weighted spectral traces (top) and a set of multi-component spectral traces (bottom) extracted from captured data for a reaction location according to specific embodiments of the invention.

[0037] FIG. 25 is a flow chart illustrating an example method of pulse detection in a dye trace according to specific embodiments of the invention.

[0038] FIG. 26 is a diagram illustrating an example of analysis of one pulse in one spectral trace according to specific embodiments of the invention.

[0039] FIG. 27 is a diagram illustrating a number of pulses associated with incorporation events and showing that some pulses with low amplitude are likely to be indicative of actual incorporation events.

[0040] FIG. 28 is a flow chart illustrating an example method of multi-timescale pulse detection according to specific embodiments of the invention.

[0041] FIG. 29 is a diagram illustrating increasing a time window and reducing a threshold for pulse detection according to specific embodiments of the invention.

[0042] FIG. 30 is a diagram illustrating algorithmic branch detection according to specific embodiments of the invention.

[0043] FIG. 31 is a flow chart illustrating an example method of a pulse merging analysis according to specific embodiments of the invention.

[0044] FIG. 32 is a flow chart illustrating an example method of pulse classification according to specific embodiments of the invention.

[0045] FIG. 33 is a diagram illustrating an example method of pulse classification according to specific embodiments of the invention.

[0046] FIG. 34 is a flow chart illustrating an example method of determining pulse metrics according to specific embodiments of the invention.

[0047] FIG. 35 is a flow chart illustrating an example method of consensus determination according to specific embodiments of the invention.

[0048] FIG. 36 is a diagram illustrating an example HMM model used in consensus calling according to specific embodiments of the invention.

[0049] FIG. 37 is a diagram illustrating an example HMM model with a branch prediction probability according to specific embodiments of the invention.

[0050] FIG. 38 is a diagram illustrating training an HMM with a set of possibly predictive error parameters and a known read sequence of a known DNA sequence template.

[0051] FIG. 39 is a diagram illustrating employing an HMM with a set of predictive error parameters, a known read sequence, and an unknown DNA sequence template.

[0052] FIG. 40 is a block diagram showing a representative example logic device in which various aspects of the present invention may be embodied.

DETAILED DESCRIPTION OF THE INVENTION

I. Sequencing by Incorporation Processes and Systems

[0053] The present invention is generally directed to novel processes, and particularly computer implemented processes, software and systems for monitoring and characterizing optical signals from analytical systems, and particularly systems that produce signals related to the sequence of nucleic acids in a target or template nucleic acid sequence, using a sequencing by incorporation process. The present invention is also generally directed to novel processes for analyzing optical and associated data from sequencing by incorporation processes to ultimately determine a nucleotide base sequence (also referred to herein as "base calling"). The present invention is also generally directed to novel processes for analyzing sequencing by incorporation processes from many reactions locations in real time.

[0054] In sequencing by incorporation methods, the identity of the sequence of nucleotides in a template nucleic acid sequence is determined by identifying each complementary base that is added to a nascent strand being synthesized against the template sequence, as such bases are added. While detection of added bases may be a result of detecting a byproduct of the synthesis or extension reaction, e.g., detecting released pyrophosphate, in many systems and processes, added bases are labeled with fluorescent dyes that permit their detection. By uniquely labeling each base with a distinguishable fluorescent dye, one attaches a distinctive detectable characteristic to each dye that is incorporated, and as a result provides a basis for identification of an incorporated base, and by extension, its complementary base upon the template sequence.

[0055] A number of sequencing by incorporation methods utilize a solid phase immobilized synthesis complex that includes a DNA polymerase enzyme, a template nucleic acid sequence, and a primer sequence that is complementary to a portion of the template sequence. The fluorescently labeled nucleotides are then added to the immobilized complex and if complementary to the next base in the template adjacent to the

primer sequence, they are incorporated onto the 5' end of the primer as an extension reaction product.

[0056] In some cases, the labeled bases are added under conditions that prevent more than a single nucleotide addition. Typically, this is accomplished through the inclusion of a removable extension terminating group on the 5' position of the added nucleotide, which prevents any further extension reactions. In some cases, the removable terminating group may include the fluorescent label. In this context, the immobilized complex is interrogated with one or more labeled nucleotide analogs. When a labeled analog is added, the extension reaction stops. The complex is then washed to eliminate all unincorporated labeled nucleotides. Incorporation is then determined based upon the presence of a particular fluorescent label with the immobilized complex, indicating incorporation of the base that was so labeled. The removable chain terminating group and the label, which in some cases may comprise the same group, are then removed from the extension product and the reaction and interrogation is repeated, stepwise, along the template sequence.

[0057] In an alternative and preferred aspect, incorporation events are detected in real-time as the bases are incorporated into the extension product. Briefly, this is accomplished by providing the complex immobilized within an optically confined space or otherwise resolved as an individual molecular complex. Nucleotide analogs that include fluorescent labels coupled to the polyphosphate chain of the analog are then exposed to the complex. Upon incorporation, the nucleotide along with its label is retained by the complex for a time and in a manner that permits its detection that is distinguishable from detection of random diffusion of unincorporated bases. Upon completion of incorporation, all but the alpha phosphate group of the nucleotide is cleaved away, liberating the label from retention by the complex, and diffusing the signal from that label. Thus, during an incorporation event, a complementary nucleotide analog including its fluorescent labels is effectively "immobilized" for a time at the incorporation site, and then the fluorescent label is subsequently released and diffuses away when incorporation is completed. According to specific embodiments of the invention, detecting the localized "pulses" of fluorescent tags at the incorporation site, and distinguishing those pulses from a variety of other signals and background noise as described below, allows the invention to effectively call bases in real-time as they are being incorporated. In conjunction with optical confinements and/or single molecule resolution techniques, the signal resulting from incorporation can have one or more of increased intensity and duration as compared to random diffusion events and/or other non-incorporation events.

[0058] In all of the foregoing aspects, optical signal data is required to be processed to indicate real incorporation events as compared to other signals derived from non-incorporation events, and to identify the bases that are incorporated in those real incorporation events. The processing requirements become even greater where multiple different colored labels are used in interrogating larger and larger numbers of immobilized complexes arrayed over reaction substrates.

[0059] For purposes of the present invention, the processes and systems will be described with reference to detection of incorporation events in a real time, sequence by incorporation process, e.g., as described in U.S. Pat. Nos. 7,056,661, 7,052,847, 7,033,764 and 7,056,676 (the full disclosures of which are incorporated herein by reference in their entirety for all purposes), when carried out in arrays of discrete reaction

regions or locations. An exemplary sequencing system for use in conjunction with the invention is shown in FIG. 1. As shown, the system 100 includes a substrate 102 that includes a plurality of discrete sources of optical signals, e.g., reaction wells or optical confinements or reaction locations 104. In typical systems, reaction locations 104 are regularly spaced and thus substrate 102 can also be understood as an array 102 of reaction locations 104. An excitation light source, e.g., laser 106, is optionally provided in the system and is positioned to direct excitation radiation at the various signal sources. This is typically done by directing excitation radiation at or through appropriate optical components, e.g., dichroic 108 and objective lens 110, that direct the excitation radiation at the substrate or array 102, and particularly the signal sources 104. Emitted signals from source 104 are then collected by the optical components, e.g., objective lens 110, and passed through additional optical elements, e.g., dichroic 108, prism 112 and lens 114, until they are directed to and impinge upon an optical detection system, e.g., detector array (or camera) 116. The signals are then detected by detector array 116, and the data from that detection is transmitted to an appropriate data processing unit, e.g., computer 118, where the data is subjected to interpretation, analysis, and ultimately presented in a user ready format, e.g., on display 120, printout 122 from printer 124, or the like, or may be stored in an appropriate database, transmitted to another computer system, or recorded onto tangible media for further analysis and/or later review. Connection of the detector to the computer may take on a variety of different forms. For example, in preferred aspects, the detector is coupled to appropriate Analog to Digital (A/D) converter that is then coupled to an appropriate connector in the computer. Such connections may be standard USB connections, Firewire® connections, Ethernet connections or other high speed data connections. In other cases, the detector or camera may be formatted to provide output in a digital format and be readily connected to the computer without any intermediate components.

[0060] This system, and other hardware descriptions herein, are provided solely as a specific example of sample handling and image capture hardware to provide a better understanding of the invention. It should be understood, however, that the present invention is directed to data analysis and interpretation of a wide variety of real-time fluorescent detecting systems, including systems that use substantially different illumination optics, systems that include different detector elements (e.g., EB-CMOS detectors, CCD's, etc.), and/or systems that localize a template sequence other than using the wave-guides described herein.

[0061] In the context of the nucleic acid sequencing methods described herein, it will be appreciated that the signal sources each represent sequencing reactions, and particularly, polymerase mediated, template dependent primer extension reactions, where in preferred aspects, each base incorporation event results in a prolonged illumination (or localization) of one of four differentially labeled nucleotides being incorporated, so as to yield a recognizable pulse that carries a distinguishable spectral profile or color.

[0062] As noted previously, the present invention is generally directed to machine or computer implemented processes, and/or software incorporated onto a computer readable medium instructing such processes, as set forth in greater detail below. As such, signal data generated by the reactions and optical systems described above, is input or otherwise received into a computer or other data processor, and sub-

jected to one or more of the various process steps or components set forth below. Once these processes are carried out, the resulting output of the computer implemented processes may be produced in a tangible or observable format, e.g., printed in a user readable report, displayed upon a computer display, or it may be stored in one or more databases for later evaluation, processing, reporting or the like, or it may be retained by the computer or transmitted to a different computer for use in configuring subsequent reactions or data processes.

[0063] Computers for use in carrying out the processes of the invention can range from personal computers such as PC or Macintosh® type computers running Intel Pentium or DuoCore processors, to workstations, laboratory equipment, or high speed servers, running UNIX, LINUX, Windows®, or other systems. Logic processing of the invention may be performed entirely by general purposes logic processors (such as CPU's) executing software and/or firmware logic instructions; or entirely by special purposes logic processing circuits (such as ASIC's) incorporated into laboratory or diagnostic systems or camera systems which may also include software or firmware elements; or by a combination of general purpose and special purpose logic circuits. Data formats for the signal data may comprise any convenient format, including digital image based data formats, such as JPEG, GIF, BMP, TIFF, or other convenient formats, while video based formats, such as avi, mpeg, mov, rmv, or other video formats may be employed. The software processes of the invention may generally be programmed in a variety of programming languages including, e.g., Matlab, C, C++, C#, .NET, Visual Basic, Python, JAVA, CGI, and the like.

[0064] While described in terms of a particular sequencing by incorporation process or system, it will be appreciated that certain aspects of the processes of the invention may be applied to a broader range of analytical sequencing or other operations and varying system configurations than those described for exemplary purposes.

[0065] FIG. 2 provides a single captured image frame of signal data on a portion of an EMCCD, where light portions correspond to optical signals that are incident upon the detector. Over time, these light areas appear as a series of flashes. Where those flashes have the appropriate characteristics, they are identified as an incorporation based signal, and assigned a color, based upon their spectral composition, e.g., as determined from their location of the detector.

II. Signal Data Processing Overview

[0066] In general, the present invention is directed to automated processes, and machine readable software that instructs such processes, for deciphering the signal data from a detection system that is optically coupled to any of the foregoing reactions, and particularly where such processes identify the incorporation of a nucleotide or nucleotide analog in a template dependent fashion, and identify the label associated with the incorporated analog and by extension, the analog and its complementary base in the template sequence.

[0067] A general flow chart illustrating the processing of signal data is provided in FIG. 3. In general, signal data is received by the processor at step 300. A number of initial calibrations operations may be applied as step 302. Some of these initial calibration steps may be performed just once at the beginning of a run or on a more continuous basis during the run. These initial calibration steps can include such things as centroid determination, alignment, gridding, drift correc-

tion, initial background subtraction, noise parameter adjustment, frame-rate adjustment, etc. Some of these initial calibration steps, such as binning, may involve communication from the processor back to the detector/camera, as discussed further below.

[0068] Generally, some type of spectral trace determination/spectral trace extraction/spectral filters are applied to the initial signal data at step 302. Some of all of this filter step may optionally be carried out at a later point in the process, e.g., after the pulse identification step 304. The spectral trace extraction/spectral filters may include a number of noise reduction and other filters as set forth elsewhere herein. Spectral trace determination is performed at this stage for many of the example systems discussed herein because the initial signal data received are the light levels, or photon counts, captured by a series of adjacent pixel detectors. For example, in one example system, 14 pixels (or intensity levels) from 14 positions are captured for an individual wave-guide at each frame. Light of different frequencies or spectrum will fall on more than one of the 14 positions and there is generally some overlap and possibly substantial overlap. According to specific embodiments of the invention, spectral trace extraction may be performed using various analysis, as discussed below, that provide the highest signal-to-noise ratio for each spectral trace.

[0069] As an alternative to a spectral trace determination, methods of the invention may also analyze a single signal derived from the intensity levels at the multiple pixel positions (this may be referred to as a summed spectral signal or a gray-scale spectral signal or an intensity level signal). In many situations, it has been found that spectral extraction, however, provides better SNR and therefore pulse detection when extracted spectral traces are analyzed for pulses somewhat separately. In further embodiments, a method according to the invention may analyze the multiple captured pixel data using a statistical model such as a Hidden Markov Model. In present systems, however, determining multiple (e.g., four) spectral traces from the initial signal data has proven a preferred method.

[0070] Whether the signal can be categorized as a significant signal pulse or event is determined at step 304. In some example systems, because of the small number of photons available for detection and because of the speed of detection, various statistical analysis techniques may be performed in determining whether a significant pulse has been detected.

[0071] If the signal is identified as a significant pulse or signal event at step 304, a further optional spectral profile comparison may be performed to verify the spectral assignment. This spectral profile comparison is optional in embodiments where spectral traces are determined prior to or during pulse identification. Once a color is assigned to a given incorporation signal, that assignment is used to call either the base incorporated, or its complement in the template sequence, at step 308. The compilation of called bases is then subjected to additional processing at step 310, to provide linear sequence information, e.g., the successive sequence of nucleotides in the template sequence, assemble sequence fragments into longer contigs, or the like.

[0072] As noted above, the signal data is input into the processing system, e.g., an appropriately programmed computer or other processor. Signal data may input directly from a detection system, e.g., for real time signal processing, or it may be input from a signal data storage file or database. In some cases, e.g., where one is seeking immediate feedback on

the performance of the detection system, adjusting detection or other experimental parameters, real-time signal processing will be employed. In some embodiments, signal data is stored from the detection system in an appropriate file or database and is subject to processing in post reaction or non-real time fashion.

[0073] The signal data used in conjunction with the present invention may be in a variety of forms. For example, the data may be numerical data representing intensity values for optical signals received at a given detector or detection point of an array based detector. Signal data may comprise image data from an imaging detector, such as a CCD, EMCCD, ICCD or CMOS sensor. In either event, signal data used according to specific embodiments of the invention generally includes both intensity level information and spectral information. In the context of separate detector elements, such spectral information will generally include identification of the location or position of the detector portion (e.g., a pixel) upon which an intensity is detected. In the context of image data, the spectral image data will typically be the data derived from the image data that correlates with the calibrated spectral image data for the imaging system and detector when the system includes spectral resolution of overall signals, e.g., as shown in FIG. 1. As discussed elsewhere herein, the spectral data may be obtained from the image data that is extracted from the detector, or alternatively, the derivation of spectral data may occur on the detector such that spectral data will be extracted from the detector.

[0074] For the sequencing methods described above, there will be a certain amount of optical signal that is detected by the detection system that is not the result of a signal from an incorporation event. Such signal, referred to hereafter as "noise" may derive from a number of sources that may be internal to the monitored reaction, internal to the detection system and/or external to all of the above. Examples of noise internal to the reaction being monitored includes, e.g.: presence of fluorescent labels that are not associated with a detection event, e.g., liberated labels, labels associated with unincorporated bases in diffused in solution, bases associated with the complex but not incorporated; presence of multiple complexes in an individual observation volume or region; non-specific adsorption of dyes or nucleotides to the substrate or enzyme complex within an observation volume; contaminated nucleotide analogs, e.g., contaminated with other fluorescent components; other reaction components that may be weakly fluorescent; spectrally shifting dye components, e.g., as a result of reaction conditions; and the like.

[0075] Sources of noise internal to the detection system, but outside of the reaction mixture can include, e.g., reflected excitation radiation that bleeds through the filtering optics; scattered excitation or fluorescent radiation from the substrate or any of the optical components; spatial cross-talk of adjacent signal sources; auto-fluorescence of any or all of the optical components of the system; read noise from the detector, e.g., CCDs, gain register noise, e.g., for EMCCD cameras, and the like. Other system derived noise contributions can come from data processing issues, such as background correction errors, focus drift errors, autofocus errors, pulse frequency resolution, alignment errors, and the like. Still other noise contributions can derive from sources outside of the overall system, including ambient light interference, dust, and the like.

[0076] These noise components contribute to the background photons underlying any signal pulses that may be

associated with an incorporation event. As such, the noise level will typically form the limit against which any signal pulses may be determined to be statistically significant.

[0077] Identification of noise contribution to overall signal data may be carried out by a number of methods, including, for example, signal monitoring in the absence of the reaction of interest, where any signal data is determined to be irrelevant. Alternatively, and preferably, a baseline signal is estimated and subtracted from the signal data that is produced by the system, so that the noise measurement is made upon and contemporaneously with the measurements on the reaction of interest. Generation and application of the baseline may be carried out by a number of means, which are described in greater detail below.

[0078] In accordance with the present invention, signal processing methods distinguish between noise, as broadly applied to all non-significant pulse based signal events, and significant signal pulses that may, with a reasonable degree of confidence, be considered to be associated with, and thus can be tentatively identified as, an incorporation event. In the context of the present invention, a signal event is first classified as to whether it constitutes a significant signal pulse based upon whether such signal event meets any of a number of different pulse criteria. Once identified or classified as a significant pulse, the signal pulse may be further assessed to determine whether the signal pulse constitutes an incorporation event and may be called as a particular incorporated base. As will be appreciated, the basis for calling a particular signal event as a significant pulse, and ultimately as an incorporation event, will be subject to a certain amount of error, based upon a variety of parameters as generally set forth herein. As such, it will be appreciated that the aspects of the invention that involve classification of signal data as a pulse, and ultimately as an incorporation event or an identified base, are subject to the same or similar errors, and such nomenclature is used for purposes of discussion and as an indication that it is expected with a certain degree of confidence that the base called is the correct base in the sequence, and not as an indication of absolute certainty that the base called is actually the base in a given position in a given sequence.

[0079] One such signal pulse criterion is the ratio of the signals associated with the signal event in question to the level of all background noise ("signal to noise ratio" or "SNR"), which provides a measure of the confidence or statistical significance with which one can classify a signal event as a significant signal pulse. In distinguishing a significant pulse signal from systematic or other noise components, the signal generally must exceed a signal threshold level in one or more of a number of metrics, including for example, signal intensity, signal duration, temporal signal pulse shape, pulse spacing, and pulse spectral characteristics.

[0080] By way of a simplified example, signal data may be input into the processing system. If the signal data exceeds a signal threshold value in one or more of signal intensity and signal duration, it may be deemed a significant pulse signal. Similarly, if additional metrics are employed as thresholds, the signal may be compared against such metrics in identifying a particular signal event as a significant pulse. As will be appreciated, this comparison will typically involve at least one of the foregoing metrics, and preferably at least two such thresholds, and in many cases three or all four of the foregoing thresholds in identifying significant pulses.

[0081] Signal threshold values, whether in terms of signal intensity, signal duration, pulse shape, spacing or pulse spec-

tral characteristics, or a combination of these, will generally be determined based upon expected signal profiles from prior experimental data, although in some cases, such thresholds may be identified from a percentage of overall signal data, where statistical evaluation indicates that such thresholding is appropriate. In particular, in some cases, a threshold signal intensity and/or signal duration may be set to exclude all but a certain fraction or percentage of the overall signal data, allowing a real-time setting of a threshold. Again, however, identification of the threshold level, in terms of percentage or absolute signal values, will generally correlate with previous experimental results. In alternative aspects, the signal thresholds may be determined in the context of a given evaluation. In particular, for example, a pulse intensity threshold may be based upon an absolute signal intensity, but such threshold would not take into account variations in signal background levels, e.g., through reagent diffusion, that might impact the threshold used, particularly in cases where the signal is relatively weak compared to the background level. As such, in certain aspects, the methods of the invention determine the background fluorescence of the particular reaction in question, including, in particular, the contribution of freely diffusing dyes or dye labeled analogs into a zero mode waveguide, and set the signal threshold above that actual background by the desired level, e.g., as a ratio of pulse intensity to background fluorophore diffusion, or by statistical methods, e.g., 5 sigma, or the like. By correcting for the actual reaction background, such as fluorophore diffusion background, the threshold is automatically calibrated against influences of variations in dye concentration, laser power, or the like. By reaction background is meant the level of background signal specifically associated with the reaction of interest and that would be expected to vary depending upon reaction conditions, as opposed to systemic contributions to background, e.g., autofluorescence of system or substrate components, laser bleedthrough, or the like.

[0082] In particularly preferred aspects that rely upon real-time detection of incorporation events, identification of a significant signal pulse may rely upon a signal profile that traverses thresholds in both signal intensity and signal duration. For example, when a signal is detected that crosses a lower intensity threshold in an increasing direction, ensuing signal data from the same set of detection elements, e.g., pixels, are monitored until the signal intensity crosses the same or a different intensity threshold in the decreasing direction. Once a peak of appropriate intensity is detected, the duration of the period during which it exceeded the intensity threshold or thresholds is compared against a duration threshold. Where a peak comprises a sufficiently intense signal of sufficient duration, it is called as a significant signal pulse. FIG. 3 provides a schematic of the process flow for an exemplary pulse classification aspect of the invention.

[0083] In addition to, or as an alternative to using the intensity and duration thresholds, pulse classification may employ a number of other signal parameters in classifying pulses as significant. Such signal parameters include, e.g., pulse shape, spectral profile of the signal, e.g., pulse spectral centroid, pulse height, pulse diffusion ratio, pulse spacing, total signal levels, and the like.

[0084] Either following or prior to identification of a significant signal pulse, signal data may be correlated to a particular signal type. In the context of the optical detection schemes used in conjunction with the invention, this typically denotes a particular spectral profile of the signal giving rise to

the signal data. In particular, the optical detection systems used in conjunction with the methods and processes of the invention are generally configured to receive optical signals that have distinguishable spectral profiles, where each spectrally distinguishable signal profile may generally be correlated to a different reaction event. In the case of nucleic acid sequencing, for example, each spectrally distinguishable signal may be correlated or indicative of a specific nucleotide incorporated or present at a given position of a nucleic acid sequence. Consequently, the detection systems include optical trains that receive such signals and separate the signals based upon their spectra. The different signals are then directed to different detectors, to different locations on a single array based detector, or are differentially imaged upon the same imaging detector (See, e.g., U.S. Patent Publication No. 2007/0036511, which is incorporated herein by reference in its entirety for all purposes).

[0085] In the case of systems that employ different detectors for different signal spectra, assignment of a signal type (for ease of discussion, referred to hereafter as "color classification" or "spectral classification") to a given signal is a matter of correlating the signal pulse with the detector from which the data derived. In particular, where each separated signal component is detected by a discrete detector, a signal's detection by that detector is indicative of the signal classifying as the requisite color.

[0086] In preferred aspects, however, the detection systems used in conjunction with the invention utilize an imaging detector upon which all or at least several of the different spectral components of the overall signal are imaged in a manner that allows distinction between different spectral components. Thus, multiple signal components are directed to the same overall detector, but may be incident upon wholly or partly different regions of the detector, e.g., imaged upon different sets of pixels in an imaging detector, and give rise to distinguishable spectral images (and associated image data). As used herein, spectra or spectral image generally indicates a pixel image or frame (optionally data reduced to one dimension) that has multiple intensities caused by the spectral spread of an optical signal received from a reaction location.

[0087] In its simplest form, it will be understood that assignment of color to a signal event incident upon a group of contiguous detection elements or pixels in the detector would be accomplished in a similar fashion as that set forth for separate detectors. In particular, the position of the group of pixels upon which the signal was imaged, and from which the signal data is derived, is indicative of the color of the signal component. In particularly preferred aspects, however, spatial separation of the signal components may not be perfect, such that signals of differing colors are imaged on overlapping sets of pixels. As such, signal identification will generally be based upon the aggregate identity of multiple pixels (or overall image of the signal component) upon which a signal was incident.

[0088] The spectral classification, or identification of a color, associated with a given signal image on a detector may be accomplished by a number of methods. In particularly preferred aspects, a spectral image associated with a given signal (which may or may not be an incorporation event signal) is compared to a standard set of spectral image profiles associated with the signal events for which the system is being interrogated. Restated, a standard set of spectral image profiles are determined for the labels associated with the four different nucleotides and/or incorporation of those nucle-

otides, and those standards are used as comparators in identifying to which color a given unknown spectral image corresponds.

[0089] In a particular exemplary calibration process, a signal source, such as a reaction region is illuminated while containing only the individual fluorescent labels or fluorescently labeled nucleotide analogs of one dye color that give rise to signals during the monitored reaction, e.g., in the absence of the reaction complex. The spectral image for each color of dye is then stored for use in the later comparison with the spectral images from actual reaction derived signals. This standard set is then used as the comparator in identifying whether the spectral image from an actual signal event can be assigned to a given color with an acceptable level of confidence, and if so, what that color is. In some cases, the spectral profiles may be determined based upon theoretical models of the optical system and the emission spectra of the signal producing reagents, e.g., labeled nucleotides, without the need for empirical determination of the standard spectral images.

[0090] As noted above, because signal separation may not be perfect in imaging signals upon a detector array, the comparison of a given signal's spectral image to the standard spectral image profiles for the various colors of signals will assess the confidence with which a color may be assigned to a given signal event, based upon a number of parameters. By way of example, whether a given spectral image is identified as matching one of the standard spectral image profiles may be determined by subjecting the comparison to any of a variety of statistical correlation evaluations including, e.g., cross-correlation tests, Ω^2 , least squares fit, and the like.

[0091] As will be appreciated, the steps of incorporation signal identification and color assignment may be performed in either order and are not dependent upon each other. Restated, one may first assign a color to the signal before categorizing it as a significant pulse, or alternatively, one may first categorize a signal as a significant pulse and then assign a color to that pulse.

[0092] Once a particular signal is identified as a significant pulse and is assigned a particular spectrum, the spectrally assigned pulse may be further assessed to determine whether the pulse can be called an incorporation event and, as a result, call the base incorporated in the nascent strand, or its complement in the template sequence. Calling of bases from color assigned pulse data will typically employ tests that again identify the confidence level with which a base is called. Typically, such tests will take into account the data environment in which a signal was received, including a number of the same data parameters used in identifying significant pulses, etc. For example, such tests may include considerations of background signal levels, adjacent pulse signal parameters (spacing, intensity, duration, etc.), spectral image resolution, and a variety of other parameters. Such data may be used to assign a score to a given base call for a color assigned signal pulse, where such scores are correlative of a probability that the base called is incorrect, e.g., 1 in 100 (99% accurate), 1 in 1000 (99.9% accurate), 1 in 10,000 (99.99% accurate), 1 in 100,000 (99.999% accurate), or even greater. Similar to PHRED or similar type scoring for chromatographically derived sequence data, such scores may be used to provide an indication of accuracy for sequencing data and/or filter out sequence information of insufficient accuracy.

[0093] Once a base is called with sufficient accuracy, subsequent bases called in the same sequencing run, and in the same primer extension reaction, may then be appended to each previously called base to provide a sequence of bases in the overall sequence of the template or nascent strand. Iterative processing and further data processing, as described in greater detail below, can be used to fill in any blanks, correct any erroneously called bases, or the like for a given sequence.

[0094] While the foregoing process generally describes the processes of the invention, additional detail is provided with reference to an exemplary sequencing process, below.

III. Exemplary Sequencing by Incorporation Process and System

[0095] The processes described above are further described in the context of a particularly preferred sequence-by-incorporation analysis using as a source of signals, a gridded array of optically confined, polymerase/template/primer complexes that are exposed to four different nucleotide analogs, e.g., A, T, G and C, that are each labeled at the terminal phosphate group of either a tri-, tetra or pentaphosphate chain of the nucleotide or nucleotide analog (e.g., as a phosphate labeled nucleoside triphosphate, tetraphosphate or pentaphosphate). In preferred aspects, the optically confined complexes are provided within the observation volumes of discrete zero mode waveguide (ZMW) cores in an arrayed format. Although described in terms of optically confined reaction regions, it will be appreciated that the methods of the invention in whole or in part may be applicable to other types of sequencing by incorporation reactions and particularly those based upon immobilized reaction complexes, and more particularly, those employing optically resolvable single molecule complexes, e.g., including a single polymerase/template/primer complex.

[0096] An example of an overall sequence process comprised of three general process categories is generally shown in FIG. 4. As shown, the process includes an initial calibration step 400, in which the system is calibrated from both an instrument standpoint and a reaction standpoint, e.g., to adjust for consistent noise sources, and calibrated for color identification, e.g., using standard dye or label sets. Once an actual sequencing data run commences, the signal data deriving from the system is subjected to initial signal processing at step 402, to identify significant signal events or pulses and to extract spectral signals from the signal data. Following the initial signal processing, classified pulses are then assessed at step 404 to classify signal data as corresponding to a given spectral signal event, and consequently as corresponding to incorporation of a particular base, or its complement.

[0097] This exemplary, overall process is schematically illustrated in greater detail in the flow chart of FIG. 5. As shown, the system is initially calibrated to provide identification of the signals associated with each discrete signal source, to identify within that signal the most relevant signal portion, and to associate spectral information with different signal profiles. This calibration process, provided in greater detail below, is indicated at step 500, and is typically precedent and supplementary to the process of signal data processing from actual sequencing runs.

[0098] Following receipt of the signal data at step 502, the signal image or movie files for a given run are converted to spectral data at step 504 by comparing the overall signal data to the spectral standards created in step 500. For example, signals received from each waveguide are converted to two

dimensional time-series or one dimensional spectral time series traces. As a result, the output of the conversion or extraction step is a series of individual movies or traces that indicate the different spectral signal components over time, e.g., as a series of n signal traces. For a typical four-color sequencing process, this will typically result in four different traces, where each trace represents the signal spectrum correlated with a different standard spectral image profile. Once the data is converted to spectrally discrete traces, the different traces are subjected to a pulse recognition or classification process at step 506. As noted above, in particularly preferred aspects, the pulse recognition process identifies significant signal pulses (e.g., pulses that meet criteria of significance for assessment to determine if they are associated with an incorporation event) in each trace, and distinguishes those from background or noise signals, e.g., those resulting from normal diffusion of unincorporated label molecules or labeled nucleotides into the observation volume, non-specific adsorption of labels or analogs within or near the observation volume, or the like. The pulse recognition process, as described in greater detail, below, identifies significant pulses based upon a number of signal characteristics as described above, including whether such signals meet signal thresholds described above (intensity, duration, temporal pulse shape, pulse spacing and spectral characteristics). Once a pulse is initially identified as significant, the time collapsed spectrum for a given significant pulse is extracted and classified at step 508 by correlating the pulse spectrum to the standard spectral image for the various signal possibilities, e.g., dye colors, by comparing the pulse spectrum to the standard spectra, based upon one or a number of different pulse metrics, as set forth elsewhere herein. For example, in this process, the statistical significance of the fit of the pulse spectral image may be calculated against those spectral images for the 4 different standard dye images, e.g., using a χ^2 test, or the like.

[0099] Once a significant pulse is correlated to a given standard dye spectrum, the pulse is then subjected to the base classification process at step 510, where the spectrum assigned pulse data is further filtered based upon one or more of a number of signal parameters, which provide a basis of classification of the signal as a particular base (also referred to herein as a base classifier). The base classifier will typically comprise an algorithm that assesses the one or more signal parameters in order to classify the particular pulse as being correlative of a given base incorporation event. By way of example, such algorithms will typically comprise a multi-parameter fit process to determine whether a spectrum assigned signal pulse corresponds to an incorporation event within a selected probability range, as described in greater detail, below.

A. Calibration

1. Gridding

[0100] As noted previously, the processes of the invention are particularly useful in processing signal data from arrays of optically confined sequencing or other optically monitored reactions. In particular, the systems and processes of the invention are particularly preferred for use with arrays of zero mode waveguides in which polymerase mediated, template directed primer extension reactions are occurring, where the addition of a nucleotide to an extending primer gives rise to a fluorescent signaling event. The signals emanating from the various signal sources on the array are then imaged onto an

imaging detector, such as a CCD, ICCD, EMCCD or CMOS based detector array. As a result, prior to running a sequencing experiment, it is typically desirable to calibrate the system to the locations of the different zero mode waveguides (or other signal sources in other processes) in the overall array, or more importantly, the position upon the detector at which the different signals from each signal source are imaged.

[0101] While in some cases, location of imaged signals could be carried out on a completed sequencing run, in many cases, it is preferred to accomplish this prior to capture of sequencing data, so that signal data associated with a given signal can be combined (e.g., binned) prior to extraction from the detector array. In particular, by identifying a collection of pixels associated with a given signal source and/or a given spectral signal from that signal source, one can collapse or combine the signal data from the different pixels prior to extracting it from the detector array. This has the result of reducing the amount of overall signal data that is required to be processed in subsequent steps.

[0102] In locating the image of the different signal sources on the detector, the array is typically illuminated so as to provide an imaged signal associated with it on the detector. In the case of an array of ZMWs (zero mode waveguides), the array is trans-illuminated through the waveguide using a reference light source. The referenced light source may be a broad band light source imaged onto the detector through a narrow band-pass filter, e.g., 543 nm, as shown in step 552 in FIG. 5 or may be an emitter of narrow-band light. Such trans-illumination provides for high signal to noise levels for the image, allowing for more accurate centroid estimation for gridding, more accurate point spread function estimation, and provides a fixed spectral reference point for each imaged waveguide.

[0103] The imaged signals are then aligned to the known spacing of image sources on the waveguide array optionally employing registration marks incorporated into the array. For example, in preferred aspects, rows of waveguides in an array will include one or more blank spaces in place of a waveguide, where the blanks will be spaced at regular, known intervals, for alignment. Other registration marks might include regularly spaced image sources that are separate from the waveguides, but are at known locations and spacing relative to the waveguides in the array to permit alignment of the image to the array. Such image sources may include apertures like waveguides, or may include fluorescent or luminescent marks that provide a signal event that can be used for alignment. The gridding step also permits the identification and calibration of the system to take into account any artifacts in a given waveguide array, e.g., blank waveguides other than registration blanks, irregularly spaced waveguides, or the like. FIG. 6 is an image of a trans-illuminated waveguide array that includes four rows of approximately 100 waveguides per row. Alignment may additionally be aided by feature recognition software that correlates signal peaks in the trans-illuminated array image with the known relative locations of waveguides on an array.

[0104] During the gridding step, the position of each waveguide image in the intra-row dimension is determined. Typically, a multi-point spread function is fit to each image to identify the image centroids in the intra-row dimension (See step 504 in FIG. 5). This is schematically illustrated in FIG. 7, which provides a close-up image of a series of trans-illuminated waveguide images in a given row, showing the identified centroids for each image. Also shown in FIG. 7 is a blank

space that is used in registration of the overall array image, as discussed above. The image positions are then used in subsequent processing to associate a given set of signal data to a given waveguide, and consequently a given extension reaction, and to permit collapsing of image data from a given location to spectral images from that location.

2. Spectral Extraction

[0105] Once the location of the imaged waveguides is determined, the range of the full imaged spectra for each waveguide are identified and this range may be communicated to detector 116 to allow binning and other data reduction operations to be performed prior to extraction from the detector. By way of example, an image of each ZMW typically has its narrowest dimension along the axis of the specific row in which the waveguide is disposed (and that is orthogonal to the axis of the spectrally separated image). For purposes of discussion, the row axis is termed the spatial axis, while the axis that runs through the elongated image of the spectral components of the waveguide is termed the spectral axis. For example, in at least one exemplary system, the spatial axis dimension of an image will fall within a 5 pixel range, while the spectral axis will typically fall within a pixel range of from about 12 pixels to about 20 or more pixels, depending upon the extent of spectral separation of the image, and the size of the image in the first instance. Thus, the pixels corresponding to the full spectral image of a given waveguide may range from 60 to 100 or more pixels in a rectangular area. These pixels that are associated with each waveguide are optionally combined (binned) by the detector 116 prior to further analysis. This combination may optionally be performed upon full image data extracted from a detector, e.g., a software process, or it may be performed within the detector, e.g., a firmware process, and output to the software process.

[0106] The software based process has a number of advantages, including, e.g.: minimizing data loss during the acquisition of the image data or movie; maximizing the signal to noise ratio of pulses based on establishing flux distribution around each waveguide in spatial dimension and noise characteristics of detector and doing appropriate weighted sum of pixel intensities; detecting and compensating for any instrument drift during movie acquisition; allowance for algorithmically distinguishing some instrument systematic artifacts such as Clock Induced Charge (CIC) noise and cosmic ray events on the CCD from the signals of interest based upon the two dimensional images being processed; the ability to estimate and potentially correct spatial cross-talk between ZMWs.

[0107] Certain disadvantages of the software process include, e.g.: a decrease in the maximum frame rate of the detector camera, as more pixels are read out from the camera, reducing the ability to detect shorter pulses; and increased instrument noise compared with firmware processes described below, resulting from read-noise that is associated with each pixel that is read out.

[0108] To maximize the signal to noise ratio (SNR) of the extracted spectrum, a weighted sum of pixels along the spatial axis of the ZMW is performed. Weights that maximize the SNR are determined by the inverse variance of each pixel. The first step is to estimate ZMW flux distribution shape in the spatial dimension for each line of ZMWs. This shape for individual ZMW signals in the trans-illumination phase is identical (governed by the instrument Point Spread Function

(PSF)). This will typically provide a good estimate of the PSF for subsequent analyses, e.g., in a sequencing movie.

[0109] In measuring the PSF in the transilluminated waveguide image typically the regular nature of the grid allows for accurate estimates in the spatial dimension by summing lines centered along a line of ZMWs. By subtracting adjacent lines an accurate local background correction can be made to leave a one dimensional intensity profile of ZMWs whose shapes are governed entirely by the instrument point spread function. The instrument PSF can be modeled (e.g., by a Gaussian or Moffat function). Fits to the one dimensional profile of a line of ZMWs may solve for all ZMW amplitudes, the ZMW spacing, and a PSF width. These fits can also solve for more parameters (e.g. a polynomial model of PSF width as a function of FOV position) in order to account for second order effects, such as variation of the optical PSF across the field of view (FOV) of the camera or variations in chip geometry.

[0110] In certain additional cases, the variance of the pixel signal intensity for a given camera is also determined. Generally for CCDs this relationship is predictable and measurable being governed by the Shot noise on the detected photons and the CCD read-noise, as well as variances from the gain register for, e.g., EMCCDs. These CCD parameters are typically estimated from the calibration data using static signal data taken at different intensities, but they can also be measured from stable pixels (pulse free) in a sequencing movie. Using the PSF estimate and the signal-variance relationship the CCD pixels are weighted-summed by their inverse variance to maximize the SNR in the collapsed spectrum. Typically, this binning or data reductions process reduces a two-dimensional pixel image for each ZMW into a one-dimensional line of pixel values. The differences in pixels along this line are due to spectral refraction as described herein. Thus, each pixel of this line is at times herein referred to as a spectral pixel of a ZMW and the line of pixels is at times herein referred to as a spectra of the ZMW.

[0111] In an alternative extraction process, the binning of spectral images from each waveguide is carried out on the detector (e.g., a camera chip in a firmware controlled process). In particular, a high resolution calibration image is taken and used to establish a map for on-chip binning in the spatial axis of the CCD in hardware to essentially read out spectra directly from the camera. In contrast to the software driven process above, this process provides benefits of: reading out fewer pixels, allowing for increased maximum frame rate, for increased sensitivity to the shorter timescale pulses; fewer pixels imaged per ZMW and therefore less instrument read-noise; and less data storage of raw output from camera. Of course, certain disadvantages of this process include, e.g.: potential for data loss during acquisition from pixels not binned; lower signal to noise ratios per pulse (if pixels are not in regime where read-noise dominates); instrument stability and/or dynamic drift correction must be done by the instrument during acquisition, rather than as a software correction; a reduced ability to distinguish instrument artifacts such as CIC noise and cosmic rays based on spatial profile; and reduced ability to account for and remediate spatial cross-talk between ZMWs, due to loss of spatial information in the image.

[0112] In this process, the location of ZMW signals is determined from a full illumination frame, and on-camera (or "on-chip") binning sums (in the spatial direction) only those CCD lines associated with a line of ZMW holes which con-

tains the majority of the signal and reads out only those lines during the actual movie acquisition. This effectively turns output of the CCD from reading out images of waveguides and waveguide arrays, to directly reading out spectra from the camera chip. As noted previously, by reducing the number of pixels output from the camera over a larger field of view, higher maximum frames rates can be maintained. The optimal binning strategy is the one that maximizes the SNR of pulses from each hole.

3. Color Calibration

[0113] During the calibration process, the system is also calibrated for the image spectra from each source. In particular, in a sequencing reaction, signals associated with each of the different incorporated bases have a distinguishable spectrum. The system used in the preferred sequencing process, e.g., as schematically illustrated in FIG. 1, through the use of a wedge prism or other similar optical train component(s), directs each spectrally different signal component from a given waveguide, to a different spot on the detector, where each type of signal from a given waveguide, is separated from the others in one dimension. In particular, a wedge prism will deflect spectrally different signal components to a greater or lesser degree, depending upon their wavelengths, and image those components on different portions of the detector (See, e.g., U.S. Patent Publication No. 2007/0036511, previously incorporated herein by reference). In order to identify a signal from a sequencing reaction as corresponding to a given label, the system must first be calibrated to the shape and location on the detector at which each spectrally different signal component from each waveguide, will be imaged, e.g., as shown at step 556 of FIG. 5.

[0114] During spectral calibration, the waveguide array is provided with a standard reference label that may include each pure dye, a pure labeled nucleotide, or another relevant pure component, e.g., a polymerase/template/primer complexed labeled nucleotide. The signal of each pure label compound is then imaged upon the detector and its location is mapped. This is repeated for each different label that is to be used in a reaction. For a typical sequencing operation, this would include the four different labels used in identifying each of the four different nucleotides. The result is a spectral template or map for the overall array and for the various different labels to be used in a sequencing operation.

[0115] In many cases, the calibration spectra will be taken at different locations on a waveguide array, e.g., from different waveguides, than analytical reads, e.g., sequencing movies. As such, the positions of the spectral images are measured relative to the reference wavelength used in the transillumination phase, above, which can then be used to correlate spectral images from different waveguide locations obtained during a sequencing movie. These spectral templates can then be aligned to the different locations on the CCD as given by the centroids of the transillumination image in the sequencing movie. Typically however the spectra as seen on the CCD are coarsely sampled and the spectral shape is sensitive to the subpixel centroid location of the ZMW within the image. To improve on estimating this shape accurately the calibration spectra are taken at multiple subpixel centroids, e.g., 0.1 pixels samplings. These can then be combined into a much higher resolution spectrum than a single image can provide. With a subpixel spectral reference centroid estimated from a ZMWs transillumination image, this high resolution spectrum can then be accurately downsampled to account for any

pixelation of the camera. In addition, due to potential distortions in the optics (e.g. coma, chromatic and spherical aberration) one may also obtain calibration spectra across the field of view of the chip. In this way, a unique spectrum is used from the calibration data for a ZMWs position, thereby accounting for spatially dependent effects that may arise from the optics.

4. Other System Calibration

[0116] In addition to the location calibration and spectral calibration, it will be appreciated that other calibration processes may also be performed. For example, in some cases, the detector is calibrated by providing for an imaging step while the shutter is closed, to ascertain and calibrate for any noise that may be deriving from the detector itself. Likewise, an overall system noise calibration step may be performed in the absence of any fluorescent or other labeling components within the waveguides in the array, to ascertain and calibrate for noise that derives from the system as a whole, e.g., auto-fluorescence of the optical train components, the array substrate, etc. (See step 558 of FIG. 5). Typically, these noise sources will be factored into one or more of the filtering steps in an overall process, e.g., subtracted from overall signal levels, or in one of the other calibration steps, e.g., in identification of the image locations and/or generation of the spectral template. The system is also typically calibrated for additional noise sources deriving from the reaction itself, e.g., resulting from nonspecific adsorption of dyes within an observation volume, presence of multiple complexes, and the like, as shown at step 560 of FIG. 5.

B. Signal Extraction to Traces

[0117] Images or movies of signal data deriving from an actual sequencing reaction is processed initially based upon the calibration of the system, as set forth above. In particular, signal data is associated with a particular signal source, e.g., waveguide, in the array based upon the positional data obtained during the calibration process. The result of the calibration process, above, is a time series of spectra for each waveguide, which is stored as an image with dimensions of the number of scans and number of spectral pixels (See, e.g., FIG. 8). The next step is to identify and classify pulse spectra from this image. Since the spectral templates for experimentally represented dyes for each ZMW are known through the above calibration process, these 2D images can be converted to 1D time series signals (one for each dye). This conversion can be achieved by a number of methods. For example, one may employ a linear matrix inversion in which the dye calibration templates are used to decompose the spectrum into the individual intensities of the pure dye spectra at each time point. Alternatively, or additionally, one may employ a weighted sum of spectral pixels using the dye templates to maximize the SNR of a pure dye sum in its trace.

[0118] In either event, the signals for each waveguide are compared to the spectral template and for each located signal source, each spectral component is then collapsed into an individual trace. In particular, the signal intensity at the image location that corresponds to a particular spectral signal from a particular signal source is plotted and/or monitored as a function of time, to provide a time resolved trace of signal activity of a given color for a given waveguide. As a result, for each located waveguide using a four color sequencing process, four different traces will be generated that reflect the

intensity of the different signal components over time. An example of trace data from four spectral traces from a single waveguide is shown in FIG. 8.

[0119] As noted above, the signal data represented in each trace is an aggregate signal of the particular pixels associated with a given spectral component of the signal. In particular, an image location may include a plurality of pixels in the detector, in order to yield the most accurate data. Rather than process signal data from each pixel in the image, the overall image can be aggregated and processed as a single data unit. Aggregation of the pixel data can be accomplished in the processor but is preferably carried out in the detector, itself, as an initial process, to minimize the amount of data created by the system and subject to further processing.

C. Pulse Recognition

[0120] Once the traces have been generated for a given waveguide, they are subjected to the pulse recognition process. The pulse recognition process is schematically illustrated in the flow chart of FIG. 9. In the initial step 900, a baseline is established for the trace. Typically, the baseline may comprise signal contributions from a number of background sources (depending on the details of the spectral and trace extraction steps). For example, such noise will typically include, e.g., global (out-of-focus) background (e.g., auto-fluorescence and large scale spatial cross-talk from the optics) and diffusion (in focus) background from the individual waveguides considered). These backgrounds are generally stable on the timescales of pulses, but still may vary slowly over longer timescales. Baseline removal comprises any number of techniques, ranging from, e.g.: a median of the trace, running lowest-percentile with bias correction, polynomial and/or exponential fits, or low-pass filtering with an FFT. Generally these methods will attempt to be robust to the presence of pulses in the trace and may actually be derived at through iterative methods that make multiple passes at identifying pulses and removing them from consideration of baseline estimation.

[0121] Other baselining functions include correction for drift or decay of overall signal levels. For example, photobleaching of organic material sometimes present on the back of the waveguide array is believed to cause decay in the level of background, and thus result in a decreasing baseline over time. This same global background decay is present on portions of the substrate at which there is no waveguide, thus allowing the traces derived from these locations to be used in combination with the two dimensional global background image to estimate the contribution of this signal to every trace/channel across the chip. This component of variability can then be subtracted from each trace and is usually very effective at removing this decay. Typically, this is carried out prior to the baselining processes described above.

[0122] As shown, each trace's baseline is established at step 900. Following establishment of the baseline the traces are subjected to noise suppression filtering to maximize pulse detection (step 902). In particularly preferred aspects, the noise filter is a 'matched filter' that has the width and shape of the pulse of interest. While pulse timescales (and thus, pulse widths) are expected to vary among different dye labeled nucleotides, the preferred filters will typically look for pulses that have a general "top-hat" shape with varying overall duration. As such, a boxcar filter that looks for a pulse of prolonged duration, e.g., from about 10 ms to 100 or more ms, provides a suitable filter. This filtering is generally performed

in the time-domain through convolution or low-pass frequency domain filtering. Other filtering techniques include: median filtering (which has the additional effect of removing short timescale pulses completely from the trace depending on the timescale used), and Savitsky-Golay filtering which tends to preserve the shape of the pulse—again depending on the parameters used in the filter).

[0123] Although described in terms of a generic filtering process across the various traces, it will be appreciated that different spectral traces may have different characteristics, and thus may be subjected to trace specific filtering protocols. For example, in some cases, a given dye labeled analog (e.g., A) may have a different pulse duration for an incorporation event than another different dye labeled analog (e.g., T). As such, the filtering process for the spectral trace corresponding to the A analog will have different filtering metrics on the longer duration pulses, than for the trace corresponding to the T analog incorporation.

[0124] In identifying pulses on a filtered trace, a number of different criteria may be used. For example, one could use absolute pulse height, either with or without normalization. Alternatively, one could identify pulses from the pulse to diffusion background ratio as a metric for identifying the pulse. In still other methods, one may use statistical significance tests to identify likely pulses over the background noise levels that exist in a given analysis. The latter method is particularly preferred as it allows for variation in potential pulse intensities, and reduces the level of false positives called from noise in the baseline.

[0125] As noted previously, a number of signal parameters may be and generally are used in pulse identification (as well as in pulse classification). For purposes of illustration, however, the process illustrate in the flow chart of FIG. 9 focuses primarily on the use of two main pulse metrics, namely pulse intensity and pulse width. As will be appreciated, the process steps at step 906 and 908 may generally include any one or more of the various pulse metric comparisons set forth elsewhere herein.

[0126] As such, following filtering, standard deviation of the baselines (noise and pulses) is determined at step 904. Preferred methods for determining the standard deviation of a trace include robust standard deviation determinations including, e.g., being based upon the median absolute difference about the baseline, a Gaussian or Poisson fit to the histogram of baselined intensities, or an iterative sigma-clip estimate in which extreme outliers are excluded. Once determined for each trace, a pulse is identified if it exceeds some preset number of standard deviations from the baseline, at step 906. The number of standard deviations that constitute a significant pulse may vary depending upon a number of factors, including, for example, the desired degree of confidence in identification or classification of significant pulses, the signal to noise ratio for the system, the amount of other noise contributions to the system, and the like. In a particularly preferred aspect, the up-threshold for an incorporation event, e.g., at the initiation of a pulse in the trace, is set at about 5 standard deviations or greater, while the down-threshold (the point at which the pulse is determined to have ended) is set at 1.25 standard deviations. The pulse width is then determined from the time between the up and down thresholds at step 910. Once significant pulses are initially identified, they are subjected to further processing to determine whether the pulse can be called as a particular base incorporation event at step 912, and as described in greater detail, below.

[0127] In some cases, multiple passes are made through traces examining pulses at different timescales, from which a list of non-redundant pulses detected at such different time thresholds may be created. This typically includes analysis of unfiltered traces in order to minimize potential pulse overlap in time, thereby maximizing sensitivity to pulses with width at or near the highest frame rate of the camera. This allows the application of pulse shape or other metrics to pulses that inherently operate on different timescale. In particular, an analysis at longer timescales may establish trends not identifiable at shorter timescales, for example, identifying multiple short timescale pulses actually correspond to a single longer, discrete pulse.

[0128] In addition, some pulses may be removed from consideration/evaluation, where they may have been identified as the result of systematic errors, such as through spatial cross-talk of adjacent waveguides, or spectral cross-talk between detection channels for a given waveguide (to the extent such issues have not been resolved in the calibration processes, supra). Typically, the calibration process will identify spectral and spatial cross-talk coefficients for each waveguide, and thus allow such components to be corrected.

[0129] Pulse recognition, e.g., on the one dimensional traces, as described above, may provide sufficient distinction to classify pulses as corresponding to particular dyes, and consequently, particular bases, based purely on their peak height. In most preferred aspects, however, the pulses identified for each waveguide are used to return to the waveguide's spectra to extract individual waveguide's spectra for each pulse for additional pulse metrics and to identify any interfering signal components, such as whether a detected pulse in a trace is due to spectral cross-talk.

D. Pulse Spectrum Extraction and Classification

[0130] Classification of an extracted pulse spectrum is then carried out by comparing the extracted spectrum to the spectra of the standard dye sets established in the calibration process. A number of comparative methods may be used to generate a comparative metric for this process. For example, in preferred aspects, a Ω^2 test is used to establish the goodness of fit of the comparison. In a particular example, for an extracted pulse spectrum (S_i), the amplitude (A) of the fit of an individual dye spectral shape, as measured from the pure dye calibration spectrum, P_i , is the only variable to solve and will have a Ω^2 value of:

$$\lambda^2 = \sum_i \frac{(S_i - AP_i)^2}{V_i}$$

[0131] The probability that the pure dye spectrum fits with the extracted spectrum is then derived from the Ω^2 probability distribution (with a number of degrees of freedom for the number of data points used, v).

[0132] The classification of a given pulse spectrum is then identified based upon calculating values for each of the four different dyes. The lowest Ω^2 value (and the highest probability fit), assigns the pulse to that particular dye spectrum, and the pulse is called as corresponding to that dye.

[0133] Again, other techniques may be employed in classifying a pulse to a particular spectrum, including for example, measuring correlation coefficients for each of the 4

possible dyes for the spectrum, with the highest correlation providing the indication to which base or dye the pulse will be classified.

[0134] In addition to comparison of the pulse spectra to the calibration spectra, a number of other pulse metrics may be employed in addition to a straight spectral comparison in classifying a pulse as correlating to a given dye/nucleotide. In particular, in addition to the spectral properties associated with a given dye, signals associated with incorporation of a given dye labeled nucleotide typically have a number of other characteristics that can be used in further confirming a given pulse classification. For example, and as alluded to above, different dye labeled nucleotides may have different characteristics such as pulse arrival time (following a prior pulse), pulse width, signal intensity or integrated counts (also referred to as pulse area), signal to noise ratio, power to noise ratio, pulse to diffusion ratio (ratio of pulse signal to the diffusion background signal in each waveguide), spectral fit (e.g., using a minimum Ω^2 test, or the like), spectrum centroid, correlation coefficient against a pulse's classified dye, time interval to end of preceding pulse, time interval to the ensuing pulse, pulse shape, polarization of the pulse, and the like.

[0135] In particularly preferred aspects, a plurality of these various pulse metrics are used in addition to the spectral comparison, in classifying a pulse to a given dye, with particularly preferred processes comparing two, three, five, 10 or more different pulse metrics in classifying a pulse to a particular dye/nucleotide.

E. Base Calling

[0136] Once the pulse spectrum is classified as corresponding to a particular dye spectrum, that correlation is then used to assign a base classification to the pulse. As noted above, the base classification or "calling" may be configured to identify directly the dye labeled base added to the extended primer sequence in the reaction, or it may be set to call the complementary base to that added (and for which the pulse spectrum best matches the dye spectrum). In either case, the output will be the assignment of a base to each recognized and classified pulse. An illustration of bases being called or assigned to different pulses is shown in FIG. 10, which shows such pulses on a collapsed timescale.

[0137] In an ideal situation, once a pulse is identified as significant and its spectrum is definitively identified, a base could simply be called on the basis of that information. However, as noted above, in typical sequencing runs, signal traces include a substantial amount of signal noise, false positive pulses, e.g., resulting from nonspecifically adsorbed analogs or dyes or the like. Accordingly, pulse classification can in many cases involve a more complex analysis. As with pulse identification, above, base classification typically relies upon a plurality of different signal characteristics in assigning a base to a particular identified significant pulse. In many cases, two, three, five, ten or more different signal characteristics may be compared in order to call a base from a given significant pulse. Such characteristics include those used in identifying significant pulses as described above, such as pulse width, pulse intensity, signal to noise ratio, power to noise ratio, integrated counts in pulse peak, shape of and distance/time to neighboring pulses, spectral signature of the pulse, pulse centroid location, and the like. Typically, such comparison will be based upon standard pattern recognition of the metrics used as compared to patterns of known base classifica-

tions, yielding base calls for the closest pattern fit between the significant pulse and the pattern of the standard base profile.

[0138] Comparison of pulse metrics against representative metrics from pulses associated with a known base identity will typically employ predictive or machine learning processes. In particular, a "training" database of "N previously solved cases" is created that includes the various metrics set forth above. From that database, a learning procedure is applied to the data in order to extract a predicting function from the data. A wide variety of learning procedures are known in the art and are readily applicable to the database of pulse metrics. These include, for example, linear/logistic regression algorithms, neural networks, kernel methods, decision trees, multivariate splines (MARS), support vector machines. Further, employing machine learned meta-algorithms for performing supervised learning, or "boosting" may be applied to any of the foregoing processes or any combinations of those. Briefly, such boosting incrementally adds to the current learned function. At every stage, a weak learner (i.e., one that yields an accuracy only slightly greater than chance) is trained with data, and that output is added to the learned function with some strength (proportional to how accurate the weak learner is). The data is then reweighted. Identifications that the current learned function has missed are then boosted in importance so that subsequent weak learners may be applied to attempt to correct the errors. Examples of boosting algorithms include, for example, AdaBoost, LPBoost, TotalBoost, and the like.

[0139] As will be appreciated, and as alluded to previously, assignment or classification of a particular pulse as incorporation of a particular base, e.g., employing the processes above, will typically be based, at least partially, on a desired probability score, e.g., probability that the called base is accurate. As noted, the probability scores for base calling, like PHRED scores for base calling in chromatographically identified bases, will typically take into account the closeness of fit of a pattern of signal metrics to a standard signal profile, based upon a plurality of different signal characteristics that include those elements described elsewhere herein, including the signal environment around a given pulse being called as a particular base, including adjacent pulses, adjacent called bases, signal background levels, pulse shape (height or intensity, width or duration, etc.), signal to noise ratios, and other signal contributors. Typically, preferred base calls will be made at greater than the 90% probability level (90% probability that the called base is correct), based upon the probability evaluation, preferably, greater than 95% probability level, more preferably greater than 99% probability, and even more preferably, greater than 99.9% or even 99.99% probability level.

[0140] The processes of the invention will typically be integrated with sequence arrangement processes for arranging and outputting the individual called bases into a linear sequence, and outputting such data to the user in any of a variety of convenient formats. Additionally, such processes will optionally verify and correct such sequence data based upon iterative sequencing of a given template, multiple sampling of overall sequence fragments through the sequencing of overlapping templates, and the like, to provide higher confidence in sequence data obtained.

[0141] A number of other filtering processes may be used in the overall evaluation of data from sequencing by incorporation reactions as discussed herein. For example, a number of

filtering processes may be employed to identify signal sources or waveguides that are yielding the highest quality level of data, e.g., resulting from a single fully functional polymerase/template/primer complex, immobilized on the bottom surface of the waveguide. These filters may rely upon a number of the metrics described above. Alternatively, these filters may employ holistic characteristics associated with a long time scale showing a large number of pulses, and determining whether the longer timescale metrics of the traces have characteristics of a typical sequence by incorporation trace, e.g., relatively regular, high confidence (based upon one or a number of relevant pulse metrics) pulses coming out over the course of the trace, yielding a "picket fence" appearance to the trace. Alternatively, additional components may be introduced to the reactants, e.g., labeling of the complexes, to facilitate their identification in the filtering process. As such, the existence of the indicator would be an initial filter to apply to any waveguide's data traces.

[0142] Although described in some detail for purposes of illustration, it will be readily appreciated that a number of variations known or appreciated by those of skill in the art may be practiced within the scope of present invention. To the extent not already expressly incorporated herein, all published references and patent documents referred to in this disclosure are incorporated herein by reference in their entirety for all purposes.

Further Example Embodiment

[0143] In order to further illustrate the invention, details are provided below regarding data collection and data analysis related to particular example sequencing systems. The details below are provided as a further example of embodiments of the invention and should not be taken to limit the invention. In some sequencing systems of interest, the relative weakness of detected signals, the levels of noise, the very small feature sizes of the reaction locations, and the speed and variation of the incorporation reaction, present challenges for signal data analysis and base-calling. These challenges are addressed by a number of novel data analysis methods and systems according to specific embodiments of the invention.

Example of Raw Data from an array of reaction locations

[0144] In order to better illustrate aspects of particular embodiments of the invention, characteristics of an example set of captured data are briefly described. An particular example system of the type illustrated in FIG. 1 can be further understood as follows. Substrate 102 is an ordered arrangement (e.g., an array) of reaction locations and/or reaction wells and/or optical confinements and/or reaction optical sources 104. Detector/camera 116 is an ordered array of optical signal collectors or detectors, such as a CCDs, EMC-CDs, or other devices able to detect optical signals and report intensity values. Such devices are well understood in the art and typically are known as able to detect light intensity incident on the detector during a given time interval (or frame) and able to output a numerical intensity value representative of the light intensity collected during a particular interval. Detector 116 will typically include an array of addressable areas, each able to make an intensity measure and data output. The separate areas are commonly referred to as pixels. Each pixel generally has an address or coordinate (e.g., x, y) and outputs one or more intensity levels for a given interval or frame. Pixels that output only one intensity level are sometimes referred to as gray-scale or monochrome pixels. A static pixel array of light intensity values (e.g., generally for one

interval) is commonly referred to as an image or frame. A time sequence of frames is referred to as a movie.

[0145] A detector/camera such as 116 may be capable of only the most basic functions necessary to capture and output intensity levels. Alternatively, a detector/camera such as 116 may include or be associated with logic circuitry able to perform various optical adjustments and/or data collection and/or data manipulation functions such as adjusting frame rate, correcting for noise and/or background, adjusting alignment or performing tracking, adjusting pixel size, combining indicated pixels prior to output, ignoring or filtering indicated pixels, etc. Thus, the raw data available from a detector 116 typically can be understood as a sequence of 2-dimensional arrays of pixel values at a particular frame rate. In an example system as in FIG. 1, the raw pixel data is mono-chrome. FIG. 2 is an example of a single frame or image of monochrome raw data captured by detector 116. In specific embodiments of the present invention, detector 116 captures and outputs 1 frame each 10 milliseconds (or 100 frames per second (fp.s.)). Data from One Reaction Location/Optical Source

[0146] While the present invention is generally directed to collection of data from many optical sources in parallel, some aspects of the invention are better understood with reference to data captured from a single optical source. In specific embodiments, the optical signal (or light) from one location 102 will pass through an optical train including a spectral spreading or refracting component such as prism 112 and lens 114. In a specific embodiment, the optical signal from one location 102 will generally be imaged on and detected by a rectangular or nearly rectangular area of pixels on detector 116. One dimension or axis (typically the longer dimension) is primarily due to spectral refraction and is herein referred to as the spectral axis. The other dimension (typically the shorter dimension) is referred to as the spatial axis. This axis is defined as the axis orthogonal to the spectral axis and is primarily due to the point source spread onto detector 116 through the optical train from location 102. In many preferred implementations the spatial axis will be reduced to one pixel or a few pixels using one or more known combination techniques, such as a point spread function (PSF) analysis. This reduction may be performed before or after data collection from the optical system as described below. In some example embodiments, the raw-data spectral dimension for one imaged location 102 is about 8 pixels to about 20 or more pixels. However, this value can vary widely as a result of the minimum size of pixels available in a detector 116 or other optical component and could feasibly range in the 100s or 1000s. In some example embodiments, the raw-data spatial dimension for one imaged location 102 is about 3 pixels to about 5 pixels. However, this value can also vary widely as a result of the minimum size of pixels available in a detector 116 or other optical component and could feasibly range in the 100s or 1000s. In some example embodiments, the frame-rate for raw captured images is about 100 frames per second (f.p.s.). However, this value could also vary widely depending on desired characteristics of the system and available computational and/or optical components.

Analysis of Arrays of Reaction Locations

[0147] In specific embodiments, data capture and data analysis according to the present invention includes many novel elements related to analyzing a large number of individual sequencing reactions located in an array of reaction locations or optical confinements. The invention, in specific

embodiments, addresses the difficulties that arise in such a system and takes advantage of the unique properties of the data arising from such a system.

[0148] FIG. 12 is a flow chart illustrating an example method of base calling from a sequencing array using a logic processing system according to specific embodiments of the invention. This figure illustrates a number of steps that are explained in greater detail below. Not all of these steps will be performed in all embodiments. In general terms, data capture and analysis involves: capturing pixels from a pixel detector for multiple reaction locations of a sequencing array (Step A1); determining correlations (gridding) between pixel sets and reaction locations (Step A2); performing one or more individualized calibrations and storing one or more individualized reaction location calibration parameters and/or data sets (Step A3); performing one or more collective array calibrations and storing one or more collective calibration parameters (Step A4); determining one or more data reduction parameters (Step A5); capturing a time sequence of pixel images from a pixel detector for multiple reaction locations during multiple incorporation reactions (Step A6); performing one or more data reductions on captured pixels (Step A7); for individual reaction locations, extracting separated dye spectral traces (Step A8); analyzing extracted dye separated traces to determine significant trace pulses and trace pulse characteristics or optionally to exclude locations with traces indicating poor quality reaction data (Step A9); using trace pulse start and end times (optionally from multiple traces) to identify a pulse in captured pixel data (Step A10); comparing a pulse in captured pixel data to one or more spectral templates to classify captured pulses as associated with a particular dye-label (Step A11); performing an initial base-calling for individual reaction locations (Step A12); examining pulse data from multiple reaction locations with overlapping sequence data and perform one or more sequence alignments using one or more of initial base calling from individual reaction locations, and probabilities associated with one or more of a classification, a false-positive pulse detection, a false-negative pulse detection (Step A13); outputting one or more consensus sequences produced by an alignment of pulse data, classification data, and probability data from multiple reaction locations, where output can include presentation to a user on a display or printout, recording sequence data on a tangible medium, or transmitting sequence data to an external computer system) (Step A14).

[0149] Analysis of sequencing-by-incorporation-reactions on an array of reaction locations according to specific embodiments of the invention is also illustrated graphically in FIG. 13. In this summary figure, data captured by a CAMERA is represented as a MOVIE, which is also a time sequence of SPECTRA. Spectral CALIBRATION templates are used to extract TRACES from the spectra. Pulses identified in the traces are then used to return to the SPECTRA data and from that data produce a temporally averaged PULSE SPECTRUM for each pulse. The Spectral CALIBRATION templates are then also used to classify PULSE SPECTRUM to a particular base. Base classifications and pulse and trace metrics are then stored or passed to other logic for further analysis.

Calibrations

[0150] Typically, various adjustments or calibrations are made in digital imaging systems both prior to and during image capture. These adjustments can include such things as

determining and correcting for background noise or various distortions caused by the optical and/or digital capture components, adjusting frame or shutter speed based on intensity levels, adjusting contrast in reported intensity levels, etc. Various such calibrations or adjustments may be made according to specific embodiments of the invention so long as the adjustments do not interfere with the data analysis as described below. Calibrations particular to specific embodiments of the present invention are described in more detail herein. Some of these calibration steps described herein may be performed periodically (such as once a week or once a day), other calibrations may be performed once at the beginning of a sequencing reaction data capture and analysis, and some calibrations are performed on a more continuous basis, throughout or at intervals during a reaction capture and analysis. These calibration steps can include such things as centroid determination, alignment, gridding, drift correction, initial background subtraction, noise parameter adjustment, spectral calibration, frame-rate adjustment, etc. Some calibration steps, such as binning, may involve communication from the processor back to the detector/camera, as discussed further below.

Gridding

[0151] An initial step in analyzing data from a system such as illustrated in FIG. 1 is determining which sets of pixels of detector 116 correspond to individual reaction locations 104. (In some implementations, this step could be addressed in the construction of the system, so that detectors and reaction locations are manufactured to have a fixed association.) Gridding, in particular embodiments, is an initial identification of particular reaction locations with particular areas of pixels in an image. According to specific embodiments of the invention, imaged signals are correlated to the known spacing of image sources on the waveguide array. Optionally, one or more registration marks incorporated into the array can be used. For example, in preferred aspects, rows of waveguides in an array will include one or more blank spaces in place of a waveguide, where the blanks will be spaced at regular, known intervals. Other registration marks might include regularly spaced image sources that are separate from the waveguides, but are at known locations and spacing relative to the waveguides in the array to permit alignment of the image to the array. Such image sources may include apertures like waveguides, or may include fluorescent or luminescent marks that provide a signal. Gridding in generally accomplished with an illuminated reference frame.

Determining Individualized Sub-Pixel Reference Centroids

[0152] According to specific embodiments of the invention, after an initial association of pixel areas in an image with particular ZMWs (gridding), an individualized reference centroid is determined and stored for each or nearly each ZMW. This centroid is determined by finding the geometric center or Gaussian center from a known spectrum, high SNR, narrow band light source that is imaged on detector 116 through generally the same optical train as sequencing reaction optical signals. With reference to FIG. 1, the illumination is directed through the partially transparent waveguides 104 and then through the optical train. Note that, while pixel address locations are generally integer values, formulas for determining a Gaussian center provide a decimal result. Using an accurate illumination through generally the exact optical path

allows determination of a subpixel reference centroid by using a selected number of decimal positions from the Gaussian center provide a decimal result. The reference sub-pixel centroid is stored for each ZMW and used as described further below. In specific embodiments, trans-illumination is provided by a light source with a narrow band-pass filter (e.g., 543 nm) or by a narrow band light source, such as an 730 nm light source. In an example embodiment, the subpixel reference centroid is rounded to a closest 0.1 interval. This sub-pixel reference centroid may then be used to estimate sub-pixel centroids for one or more detection signals as described herein. FIG. 14 is a flow chart illustrating an example method of determining a reference spectral centroid for a reaction location image according to specific embodiments of the invention. FIG. 15 is a diagram illustrating an initial gridding step and corrected centroids according to specific embodiments of the invention. In this figure, the central white line down the center of pixel 277 indicates centroid locations as determined by an initial gridding step. The individual dots to the left of this line indicate individually determined centroids for each ZMW as discussed above. The spline through these individual dots indicates an optionally spline fitting or similar step that can be used to smooth centroid locations and provide a smoothed individualized sub-pixel centroid.

Alignment Across Multiple Reaction Locations

[0153] FIG. 16 is a flow chart illustrating an example method of determining an alignment for reaction locations and a central axis for individual reaction locations from multiple reaction location images according to specific embodiments of the invention. After the initial gridding step and determination of a sub-pixel centroid effectively in the spatial axis, an alignment along a spatial axis may be performed to more accurately determine a location of a sub-pixel centroid in the spectral axis. This generally involves capturing pixels from a plurality of ZMWs (Step C1); determining a sub-pixel centroid of at least two ZMWs (Step C2); determining a centroid line through centroids of at least two ZMWs (Step C3); optionally, determining if the centroid line is optimally parallel to pixel rows of a pixel detector (Step C4); optionally, adjusting alignment (optically or physically) between a detector and ZMWs so that a centroid line between multiple ZMWs is optimally parallel to pixel rows in a pixel detector (Step C5); optionally, storing correction values for ZMWs with a centroid substantially off-center of a pixel (Step C6); storing an indication of a plurality of central axes for a plurality of ZMWs (Step C7); optionally transmitting an indication of binning parameters for each ZMW to the pixel detector (Step C8).

Individualized Spectral Templates and Spectral Calibration

[0154] FIG. 17 is a flow chart illustrating an example method of determining high-resolution dye spectral templates and down-sampled individualized spectral templates according to specific embodiments of the invention. In order to provide a more accurate spectral trace extraction and spectral classification, a high resolution spectral template is optionally produced using multiple ZMWs. During spectral calibration, the waveguide array is provided with a standard reference label that may include each pure dye, a pure labeled nucleotide, or another relevant pure component, e.g., a polymerase/template/primer-complexed labeled nucleotide. The signal of each pure label compound is then imaged upon the

detector using an optical train and its location is mapped. This is repeated for each different label that is to be used in a reaction. The result is a spectral template or map for the various different labels to be used in a sequencing operation. Calibration spectra may be taken at different locations on a waveguide array than analytical reads and are then aligned to the different locations on the detector using the centroids of the transillumination image. Typically however the spectra as seen on the detector are coarsely sampled and the spectral shape is sensitive to the subpixel centroid location of the ZMW within the image.

Stitching

[0155] To improve on estimating this shape accurately the calibration spectra are taken at multiple subpixel centroids, e.g., 0.1 pixels samplings. These can then be combined into a much higher resolution spectrum than a single image can provide. With a subpixel spectral reference centroid estimated from a ZMWs transillumination image, this high resolution spectrum can then be accurately downsampled to account for pixelation of the camera. In addition, due to potential distortions in the optics (e.g. coma, chromatic and spherical aberration) one may also obtain calibration spectra across the field of view of the chip. In this way, unique spectral templates are used from the calibration data for a ZMWs position, thereby accounting for spatially dependent effects that may arise from the optics. In a specific embodiment, an array of ZMWs is mechanically or optically tilted slightly during spectral calibration data collection. The tilting provides a mechanism for capturing calibration spectra with sub-pixel offsets for averaging to provide spectral templates of higher resolution. FIG. 18 is a diagram illustrating stitching high-resolution dye spectral templates from multiple captured spectral calibration data according to specific embodiments of the invention.

[0156] One method of accomplishing high resolution calibrations relies on the alignment of reaction locations (e.g., ZMWs) according to specific embodiments of the invention. To understand this, consider a vertical row of 100 ZMWs, each 10 pixels wide along the spectral axis and effectively 1 pixel high in the spatial dimension (although more spatial pixels could be used.). One means of averaging a spectral calibration template of the 100 ZMWs would be to simply average each of the 10 pixel locations separately down the 100 ZMWs. This would provide an averaged spectral template, but not a higher resolution one. Now, however, imagine tilting the vertical row of ZMWs so that the top most ZMW (designated herein as ZMW₁₀₀) was roughly one pixel tilted (e.g., for purpose of discussion, to the right, towards pixels) from the bottom most ZMW, ZMW₁. In this case, averages can be taken at sub-pixel locations, where, for example, the leftmost sub-pixel location, pixel_{0.1}, will be the average of pixel₁ for the 10 lowest ZMWs for that location (e.g., ZMW₁ through ZMW₁₀). The next sub-pixel, pixel_{0.2}, will be the average of a sliding window of 10 ZMWs (e.g., ZMW₂ through ZMW₁₁ for pixel_{0.2}, ZMW₃ through ZMW₁₂ for pixel_{0.3}, etc.). In this example, pixel_{1.1} will be an average of pixel₁ for 9 ZMWs (e.g., ZMW₉₁ through ZMW₉₉) and the second pixel for ZMW₁₀₀. In this way, in this example, 100 ZMWs with 10 points of spectral pixel resolution are averaged into one spectral template with 100 points of spectral pixel resolution.

[0157] In practice, the spectra taken for ZMW₁-ZMW₁₀₀ will in general have their spectral reference wavelength

placed at varying sub-pixel shifts, relative to the centroid pixel of the ZMW, because the centroids of the ZMWs will vary more or less uniformly across ~ 1 pixel. Now, assume that a ZMW centroid can be localized to within $\sim 1/10$ pixel as described herein. The ZMW spectra can therefore be characterized as arising from a re-binning of spectra at $10\times$ the resolution, where the higher resolution bin offset is known. The high-resolution spectrum can then be estimated by placing each ZMW spectrum in its corresponding high-resolution bin locations (shift and pitch), and then averaging the values in each bin.

Alternative Spectral Calibration Using Known Segments Added to a Template

[0158] In specific embodiments, the generation of high resolution spectral calibration templates is done periodically, such as once a day or once a week, as it involves generally four different rounds of exposing an array 102 to four different dyes, with the overhead of preparation of the array for each of the four different reactions. The high resolution spectral templates are then individualized (and optionally down-sampled) using a spectral subpixel centroid, generally during each run. In alternative embodiments, however, a spectral template can be determined for each ZMW for each sequencing reaction by including a series of known bases in a known sequence. In such a case, spectral calibration data for each dye is collected for each ZMW and averaged to provide an individualized ZMW spectral template, optionally using additional relevant data as provided herein.

Determining Background Noise

[0159] FIG. 19 is a flow chart illustrating an example method of determining background noise for reaction location images according to specific embodiments of the invention.

Software Spatial Data Reduction

[0160] FIG. 20 is a flow chart illustrating an example method of data reduction for a reaction location image according to specific embodiments of the invention. In specific embodiments, to maximize the signal to noise ratio (SNR) of the extracted spectrum, a weighted sum of pixels along the spatial axis of a ZMW (or reaction location) image area is performed. Weights that maximize the SNR are determined by the inverse variance of each pixel. In an grid array 102 of many reaction locations, optical signal detected from multiple locations can be used to determine a line of pixels through multiple locations, and this line of pixels can be used as the central line for a PSF function. By subtracting adjacent lines an accurate local background correction can be made to leave a one dimensional intensity profile of ZMWs whose shapes are governed entirely by the instrument point spread function. The instrument PSF can be modeled (e.g., by a Gaussian or Moffat function). Fits to the one dimensional profile of a line of ZMWs may solve for all ZMW amplitudes,

the ZMW spacing, and a PSF width. These fits can also solve for more parameters (e.g. a polynomial model of PSF width as a function of FOV position) in order to account for second order effects, such as variation of the optical PSF across the field of view (FOV) of the camera or variations in chip (e.g., array substrate 102) geometry.

On-Camera Data Reduction (Binning)

[0161] In an alternative extraction process, binning to derive spectral images from each location 104 waveguide is carried out on the camera chip (in a firmware controlled process). In this process, the location of ZMW signals is determined from a full illumination frame, and on-camera (or "on-chip") binning sums (in the spatial direction) only those CCD lines associated with a line of ZMW holes which contains the majority of the signal and reads out only those lines during the actual movie acquisition. This effectively turns output of the CCD from reading out images of waveguides and waveguide arrays, to directly reading out spectral-spread images from the camera chip. As noted previously, by reducing the number of pixels output from the camera over a larger field of view, higher maximum frames rates can be maintained. The optimal binning strategy is the one that maximizes the SNR of pulses from each reaction location.

Spectral Trace Extraction

[0162] After the calibration and data reduction steps described above, according to specific embodiments of the invention, the optical signal data for an individual ZMW or location 104 is a sequence or movie of a small area (e.g., an area of about 1×14 pixels) of monochrome spectral-images also referred to herein as spectra. The optical signal data for an individual ZMW can also be understood or represented as a time series of arrays (or vectors) of intensity values (e.g., a sequence of 1 by 14 intensity values). According to specific embodiments of the invention, from these intensity values, one or more spectral traces are extracted for further analysis, e.g., pulse detection. A spectral trace, as used herein, is a time series of generally a single intensity value.

[0163] FIG. 21 illustrates an example of four superimposed spectral images representing different dyes imaged from two different reaction locations according to specific embodiments of the invention. FIG. 21 provides some illustration of the nature of the spectral-spread position data and the difficulties in extracting spectral traces that are overcome according to specific embodiments of the invention. The top portion shows four superimposed graphs of 14 pixel values. Each of the four superimposed graphs represents "training data" or "calibration data" from four different known dye spectra at a particular ZMW. Represented numerically, the data at the top portion would approximately be as shown in the table below. For ease of reading, the intensity values shown in the figure are multiplied by 100 in the table below.

TABLE 1

Dye Spectra	Pixel#													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
488 nm (left most peak)	0	1	3	2	1	1	0.5	0	0	0	0	0	0	0
555 nm (next peak)	0	0	0	0	1	3	3	1	0.5	0	0	0	0	0

TABLE 1-continued

Dye Spectra	Pixel#													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
568 nm (next peak)	0	0	0	0	0.5	1	4	2	1	0.5	0	0	0	0
647 nm (right most peak)	0	0	0	0	0	0	0	0	2	5	3	10	5	0

[0164] The top graph in FIG. 21 indicates one difficulty that is addressed according to specific embodiments of the invention. The figure indicates that even in a situation of "best case" training data, there is can be substantial overlap between the pixel data captured from different dyes, e.g., the 555 and 568 nm dye spectra. In this example, this is to be expected due to the closeness of the wavelengths of the fluorescent signals. However, choice of dye may be constrained for a variety of reasons, and extraction of even closer spectra may be desired in some situations. The bottom graph illustrates another complexity. This figure represents the same data as above, but captured at a different ZMW location. As a result of slight differences in alignment of the true sub-pixel centroid of the ZMW image with the centroid pixel as determined on the detector and/or other slight differences in the optical system, there is variation in the expected data. FIG. 22 is a flow chart illustrating an example method of dye trace extraction from a time series of a reaction location images according to specific embodiments of the invention. The method involves: at a captured frame, performing a flux matching to a first dye spectrum template (Step G1); optionally, using a down-sampled high resolution dye spectral template optionally individualized to a subpixel-centroid (Step G2); storing outputting an intensity value from the flux analysis as the frame intensity for the dye (Step G3); repeating the spectral template comparison on the same captured frame compared for each dye spectrum template (Step G4); repeating the above steps at successive frames to generate one or more dye spectral traces for each ZMW (Step G5); optionally, repeat all above steps using a multi-component flux matching to generate a second set of trace intensities corrected for cross talk (Step G6). Details of this method are described further below.

[0165] FIG. 23 is a diagram illustrating using an individualized set of spectral templates for a ZMW to extract four spectral traces using a flux calculation according to specific embodiments of the invention. In a typical embodiment, four spectral traces will be extracted. In an alternative embodiment, one or more of the spectral values may be combined (such as 555 and 568 as indicated above) to derive fewer traces. In other cases, one or more additional traces may be derived. Extraction of the four spectral intensity values (or spectral pixel values) from the 14 raw values is accomplished according to specific embodiments of the invention as follows. With individualized spectral templates determined for each dye for each ZMW, trace extraction can be achieved by a number of methods, as described herein. After spectral extraction, the signal intensity at the image location that corresponds to a particular spectral signal from a particular signal source is plotted and/or analyzed as a function of time. Various techniques for spectral extraction from data similar to the spectral-spread image data are known and can be used for spectral extraction according to specific embodiments of the invention. Horne (1986), for example, discusses a number of spectral extraction techniques used in CCD-based astronomi-

cal spectroscopy. In one embodiment, spectral extraction is performed for the four intensities using the reference spectra, and spectral-spread image, according to the following equation:

$$F = \frac{\sum_i (R_i - B_i)(P_i / V_i)}{\sum_i P_i^2 / V_i}$$

where F is F is the summed flux at each spectral pixel that maximizes S/N. This technique is used for spectral extraction according to specific embodiments of the invention.

[0166] As an alternative to a spectral trace determination, methods of the invention may also analyze a single signal derived from the intensity levels at the multiple pixel positions (this may be referred to as a summed spectral signal or a gray-scale spectral signal or an intensity level signal). In many situations, it has been found that spectral extraction, however, provides better SNR and therefore pulse detection when extracted spectral traces are analyzed for pulses separately. In further embodiments, a method according to the invention may analyze the multiple captured pixel data using a statistical model such as a Hidden Markov Model. In present systems, however, determining multiple (e.g., four) spectral traces from the initial signal data has proven a preferred method.

[0167] FIG. 24 is a diagram illustrating a set of dye-weighted spectral traces (to) and a set of multi-component spectral traces (bottom) extracted from captured data for a reaction location according to specific embodiments of the invention.

Pulse Recognition

[0168] FIG. 25 is a flow chart illustrating an example method of pulse detection in a dye trace according to specific embodiments of the invention. FIG. 26 is a diagram illustrating an example of analysis of one pulse in one spectral trace according to specific embodiments of the invention.

[0169] Further refinements to a pulse detection algorithm from that described above according to specific embodiments of the invention may be made by consideration of variability in captured spectral intensity data that is due solely to the expected kinetic behavior of the underlying sequencing reaction. Such data can, for example, include noiseless traces created by a kinetic model simulator. FIG. 27 is a diagram illustrating a number of pulses associated with incorporation events and showing that some pulses with low amplitude are likely to be indicative of actual incorporation events. The figure illustrates an example of a captured data frames for a number of zero mode waveguides according to specific embodiments of the invention. It has been determined from this analysis that even in the absence of noise or other system

variability, the pulse width or duration of an actual incorporation event, as well as the between pulse intervals, stochastically varies. The variation in incorporation event timing affects the both the duration and amplitude of pulses. The duration affect is clearly understood. Amplitude is affected because of the nature of the detectors 116, which provide an intensity value in proportion to the total optical signal energy received during the collection interval. In the discussion herein, this interval is referred to as a frame.

[0170] In understanding the data shown in FIG. 27, assume that an incorporation event produces an optical signal intensity at a normalized intensity value of 1 during each capture interval (e.g., frame). Further, assume that a frame is 10 milliseconds. Then, an incorporation event that happens to begin at a frame boundary, and that happens to last for one frame, will produce a pulse with a 10 millisecond (1 frame) width and an amplitude of 1. However, it is equally likely that a 10 millisecond incorporation event will begin in the middle of one frame and complete in the middle of the next frame. In such a case, the same incorporation event will produce a pulse with a 20 millisecond (2 frame) width and an amplitude of 0.5. Furthermore, there may be a relatively high probability that an incorporation event will last only 5 milliseconds (e.g., $\frac{1}{2}$ a frame). If such an incorporation event occurs entirely during a single frame, such an incorporation event will produce a 1 frame pulse (1 frame being the shortest pulse width detectable) with an amplitude of 0.5. There is also a probability that a 5 millisecond pulse will occur exactly across 2 frames. In such a case, the same incorporation event will produce a pulse with a 20 millisecond (2 frame) width and an amplitude of 0.25. With this understanding, it will be understood that the lower amplitude pulses circled in FIG. 27 can have an equal probability of being valid significant pulses as the higher amplitude pulses.

[0171] Using the data as described above, a pulse detection algorithm according to specific embodiments of the invention, can assign confidence levels and make adjustments to account for stochastic false positive (FP) rates (e.g., detecting a pulse as a result of noise alone), stochastic false negative (FN) rates (e.g., failing to detect a pulse because it is masked by background noise), and miss-match (MM) errors (e.g., incorrectly classifying the spectra of a pulse due to its detected width and intensity.) Using such an analysis, for example, the invention can determine that optimal pulse calling thresholds are around 3.5 sigma for each channel based on the kinetic parameters of the incorporation reaction and the frame capture parameters.

[0172] According to specific embodiments of the invention, such an analysis shows that increasing the frame rate can increase the SNR of sub-frame pulse-width (PW) pulses by roughly $\sqrt{\text{fps}}$, and therefore while the FN rate is initially reduced with an increased FPS, it increases again as pulse peak SNR degrades at inverse $\sqrt{\text{fps}}$. Detection of one significant pulse as two (referred to as algorithmic branching) also is found to be an increasing problem at higher frame rates.

[0173] According to specific embodiments of the invention, an pulse detection algorithm is optimized to reduce FN rates while allowing an increase in fps rate. Such a pulse detection optimization runs the initial pulse detection algorithm using a standard "1 frame" pulse detector to generate a first pulse list. Then, an $N=2$ frame boxcar filter is applied, noise reduction is by $\sqrt{\text{QRT}(2)}$, the threshold is moved accordingly, and the pulse detector is run to find additional

detected pulses. This step can be repeated for $N=3, 4, \dots$ until N approaches or exceeds the longest expected detectable pulse. FIG. 29 is a diagram illustrating increasing a time window and reducing a threshold for pulse detection according to specific embodiments of the invention. In this example figure, a pulse between frame 23 and 30, with its highest amplitude at just over 75 counts is not detected at $N=1$ (top) due to a noise threshold set at 77 counts, resulting in a false negative (FN). With $N=2$, and a subsequent reduction of the noise threshold to about 63, the pulse is detected.

[0174] In another aspect, it has been determined that lower amplitude pulses are more prone to algorithmic branching near the noise threshold. To address this, a pulse detection according to specific embodiments of the invention uses merge heuristics as follows. For consecutive pulses of same base, for each pair estimate a mean pulse height and inter-pulse baseline. Using a noise model and/or actual noise data, estimate the statistical significance of inter-pulse region relative to peak height and merge consecutive peaks the statistical significance is <3 standard deviations. FIG. 30 illustrates two consecutive pulses between frames 20 and 34 that would be merged as discussed above according to specific embodiments of the invention, thus eliminating one FP that would occur without the merging. FIG. 31 is a flow chart illustrating an example method of a pulse merging analysis according to specific embodiments of the invention.

Pulse Recognition Using DBR (Diffusional Background Ratio)

[0175] As an alternative to pulse recognition using pulse intensity as the primary signal value as described above, according to further specific embodiments, the invention uses the ratio of pulse height to diffusion background, or DBR (diffusional background ratio) in determining significant pulses. While at times this may be referred to a component of SNR (signal to noise ratio), the term DBR is used to avoid confusion with more traditional usages of SNR. In specific embodiments, it has been found that calling pulses by intensity or by sigma can either fail due to laser intensity variations or due to background noise variations (respectively). Calling pulses by the DBR according to specific embodiments of the invention makes the "intensity" component of the pulse less variable even in the presence of laser excitation variations and regardless of background noise envelope fluctuations (due to concentration variations, extremely sticky ZMWs, etc). Furthermore, setting the pulse intensity threshold according to the intensity contributions of freely diffusing fluorophores in the ZMW provides a theoretical framework for locating a single molecule event in a ZMW and provides some immunity from other sources of signal variations and error. There are several methods of obtaining an estimate of the DBR intensity per ZMW.

[0176] Thus, in specific embodiments, pulse intensity is described not in absolute counts measured against a threshold, but as a ratio against the background diffusion of fluorophores in and above an individual ZMW.

[0177] According to specific embodiments of the invention, for any given pulse, define its DBR as: $\text{DBR} = (\text{Intensity} - \text{dcOffset}) / (\text{dcOffset} - \text{NDB})$ where $(\text{Intensity} - \text{dcOffset})$ is the average intensity of a pulse above baseline, and "NDB" is the portion of the baseline (dcOffset) that is not diffusion background (e.g., baseline from autofluorescence, base clamping, etc.). In particular embodiments, the NDB is determined from a sample movie of the array (or a neutral substrate, such as a

solid aluminum film) with the same laser and camera conditions, which provides values for NDB(ZMW).

[0178] The DBR method of pulse calling provides additional information about where in the ZMW a particular pulse originated. This information is used in specific embodiments to determine if multiples polymerase are sequencing in a ZMW, in which case data from that specific ZMW may be excluded from further data analysis. The location of a fluorophore within a ZMW can also be used as one of the parameters in the data analysis as described herein.

[0179] The maximum values of the DBR of pulses from a single ZMW also allows estimation of the ZMWs effective diameter according to specific embodiments of the invention. In a particular example implementation, this method was used to estimate the ZMW diameters in an array to vary within 13 nm.

[0180] DBR thresholding in some embodiments may be vulnerable to diameter variations of the ZMWs themselves across the array (because more diffusion will occur into larger diameter ZMWs. In specific embodiments, this is accounted for on a per-ZMW basis, for example by transmission light analysis prior to sequencing. With the size of each ZMW known or accurately analyzed, the DBR method is generally preferable to sigma-calling or intensity-calling of pulses.

Trace or Reaction Location Rejection

[0181] According to specific embodiments of the invention, as described herein, analysis of individual ZMWs includes repeated evaluation of whether a ZMW should be excluded from further analysis. Because large numbers of reaction locations are being prepared and monitored, it is expected that in some systems some percentage of reaction locations will not provide useful data. This may occur if no reaction enzyme becomes located in a particular ZMW, if more than one reaction enzyme is located in a ZMW or if a reaction enzyme is otherwise producing problematic data. Rejection of particular reaction location data streams may be performed at multiple points during the analysis where the captured data does not match expected data criteria.

Pulse Classification/Confirmation/Base Calling

[0182] FIG. 32 is a flow chart illustrating an example method of pulse classification according to specific embodiments of the invention. In this example, for each dye trace for a ZMW, the invention retrieves pulse start and end times (Step K1); these times are combined to determine a plurality of pulses in the pre-extraction captured data (Step K2); optionally, for each pulse, determine proximate background correction values from temporally close captured pixel values that are not within a pulse (Step K3); optionally, for each pulse, examine multi-component traces to avoid false pulse combinations due to dye cross-talk (Step K4); for each pulse, temporally combine captured pixel values into a temporally averaged pulse frame (Step K5); for each pulse, compare flux values to a plurality of dye spectral templates to determine a best match, optionally using proximate background correction values (Step K6); store or output the best match as the dye classification for a significant pulse (Step K7); optionally store or output a match probability for a significant pulse, optionally using proximate background correction values (Step K8); optionally store or output alternative less-probable classifications and corresponding match probabilities for a significant pulse (Step K9); optionally store or output one or

more additional pulse metrics, including inter-pulse metrics for possible use in base calling) (Step K10).

[0183] A number of comparative methods may be used to generate a comparative metric for this process. For example, in preferred aspects, a test is used to establish the goodness of fit of the comparison. In a particular example, for an extracted pulse spectrum (S_i), the amplitude (A) of the fit of an individual dye spectral shape, as measured from the pure dye calibration spectrum, P_n , is the only variable to solve and will have a Ω^2 value of:

$$\chi^2 = \sum_i \frac{(S_i - AP_i)^2}{V_i}$$

[0184] The probability that the pure dye spectrum fits with the extracted spectrum is then derived from the Ω^2 probability distribution (with a number of degrees of freedom for the number of data points used, v). The classification of a given pulse spectrum is then identified based upon calculating values for each of the four different dyes. The lowest Ω^2 value (and the highest probability fit), assigns the pulse to that particular dye spectrum, and the pulse is called as corresponding to that dye.

[0185] Again, other techniques may be employed in classifying a pulse to a particular spectrum, including for example, measuring correlation coefficients for each of the 4 possible dyes for the spectrum, with the highest correlation providing the indication to which base or dye the pulse will be classified.

[0186] In addition to comparison of the pulse spectra to the calibration spectra, a number of other pulse metrics may be employed in classifying a pulse as correlating to a given dye/nucleotide. In particular, in addition to the spectral properties associated with a given dye, signals associated with incorporation of a given dye labeled nucleotide typically have a number of other characteristics that can be used in further confirming a given pulse classification. For example, and as alluded to above, different dye labeled nucleotides may have different characteristics such as pulse arrival time (following a prior pulse), pulse width, signal intensity or integrated counts (also referred to as pulse area), signal to noise ratio, power to noise ratio, pulse to diffusion ratio (ratio of pulse signal to the diffusion background signal in each waveguide), spectral fit (e.g., using a minimum Ω^2 test, or the like), spectrum centroid, correlation coefficient against a pulse's classified dye, time interval to end of preceding pulse, time interval to the ensuing pulse, pulse shape, polarization of the pulse, and the like.

[0187] In particularly preferred aspects, a plurality of these various pulse metrics are used in addition to the spectral comparison, in classifying a pulse to a given dye, with particularly preferred processes comparing two, three, five, 10 or more different pulse metrics in classifying a pulse to a particular dye/nucleotide.

Optional Additional Trace Extraction to Avoid Conflation

[0188] As discussed herein, extraction from spectra to multiple spectral traces is may be performed according to an algorithm that maximizes the flux in each trace. As a result of this and of the fact that selected spectral dyes may have substantial overlap, in certain situations this approach will result in single incorporation pulse being detected in two

traces. Because traces are used generally to determine start and end times from the captured data, this situation does not present a problem in most cases.

[0189] However, in some cases this may cause a merging of pulses that should be associated with two differently classified incorporation events. To address, in specific embodiments of the invention, a secondary spectral trace extraction is performed that attempts to increase separation between spectral template matches. This secondary trace extraction is then used to confirm that start and end times of pulses represent a pulse in one spectral color and not in two overlapping colors. Consensus Generation and Sequence Alignment using Statistical Models and Pulse Features Background

[0190] Various techniques for automated "smart" base calling of Electrophoretic DNA sequencing data have been discussed. Electrophoretic DNA sequencing often involves trace data from four different dyes that are used to label four bases. PHRED is a base-calling program for automated sequencer traces that outputs at each base generally one of five base identifiers (A C T G and N for not identifiable) and often a quality score for each base. In PHRED processing of DNA traces, predicted peak locations in terms of migration times are determined, observed peaks are identified in the trace and are matched to predicted peak locations, sometimes omitting some peaks and splitting. Unmatched observed peaks may be checked for any peak that appears to represent a base but could not be assigned to a predicted peak in the third phase and if found, the corresponding base is inserted into the read sequence. Peaks in a PHRED analysis may be difficult to distinguish in regions where the peaks are not well resolved, noisy, or displaced (as in compressions). The PHRED algorithm typically assigns quality values to the bases, and writes the base calls and quality values to output files. PHRED can evaluate the trace surrounding each called base using four or five quality value parameters to quantify the trace quality. PHRED can use dye chemistry parameter data to do such tasks as identifying loop/stem sequence motifs that tend to result in CC and GG merged peak compressions. PHRAP is a sequence assembly program often used together with PHRED. PHRAP uses PHRED quality scores to determine highly accurate consensus sequences and to estimate the quality of the consensus sequences. PHRAP also uses PHRED quality scores to estimate whether discrepancies between two overlapping sequences are more likely to arise from random errors, or from different copies of a repeated sequence. Various expert analysis and similar systems have been proposed for analyzing such data, See, for example, U.S. Pat. No. 6,236,944, Expert system for analysis of DNA sequencing electropherograms. The use of statistical models, such as hidden Markov models (HMMs), for DNA sequencing has been discussed by several authors (See, e.g., Petros Boufounos, Sameh El-Difrawy, Dan Ehrlich, HIDDEN MARKOV MODELS FOR DNA SEQUENCING, Journal of the Franklin Institute, Volume 341, Issues 1-2, January-March 2004, Pages 23-36 Genomics, Signal Processing, and Statistics. HMMs have been discussed as an approach to DNA basecalling, using techniques such as modeling state emission densities using Artificial Neural Networks, and a modified Baum-Welch re-estimation procedure to perform training. Consensus sequences have been proposed to label training data to minimizing the need for hand-labeling.

[0191] In further specific embodiments, software methods of the present invention include techniques for generating consensus DNA sequence information of high accuracy from

a collection of less accurate reads generated by a real-time sequencing by incorporation system. In specific embodiments, two features of data typical of some such systems that motivate these techniques are: (1) the errors in the raw data are mostly insertions or deletions of base symbols from the correct sequence, rather than 'mismatches' or misidentified bases; (2) a relatively large number (e.g., 1000 or more) of data points are collected in real time for each base symbol in the raw read.

[0192] As described above, a signal intensity and signal spectrum is measured through time. This results in a large collection of data features associated with each base in the raw read sequence. The time series data are summarized by finding regions of high signal intensity 'pulses', and measuring a series of features of those pulses, such as their duration, average intensity, average spectrum, time until the following pulse, and best reference spectra match. Observable pulses are generated when nucleotides are productively incorporated by the polymerase ("incorporation pulses"), as well as by interfering processes (such as incorrect bases that stick temporarily but are not incorporated or correct bases that become illuminated temporarily, but are not fully incorporated, and then are incorporated and produce a second pulse (branching)) that introduce errors into the observed sequence. (Pulses that are entirely due to random noise may also be detected and are attempted to be identified as described above). The statistical nature of the process that generates the pulses results in wide, but measurable distributions of pulse features. Processes that generate spurious pulses generate pulses with different distributions of pulse features, although the distributions between spurious pulses and incorporation pulses will overlap.

[0193] Thus, in specific embodiments, a predictive HMM observation distribution model, is extended to not only identity of the called base, but also the features of the associated pulse. In this method each class of microscopic event (true incorporations as well as interfering events) generates pulses with different but overlapping probability distributions in the space of pulse features. The distribution over pulse feature space for each pulse type is learned from experimental data and used to generate an approximate observation distribution via density estimation techniques. In a final sequence alignment step, a most likely template (sequence) is discovered by constructing a series of trial models that maximize the likelihood of the observed data under the model, via an expectation maximization procedure.

[0194] The following example describes this algorithmically. Let $b_{ij}(O)$ represent the probability distribution of observations received when transitioning from state i to state j of the HMM. In typical alignment applications, the alphabet of observations that it is possible to observe is limited to {A,G,C,T}. That is, each pulse observed is summarized only by its base identity. When combining multiple reads of the same DNA into a single consensus read, it is necessary to resolve the ambiguity that exists between reads. For example, in a base detection from two sources as follows:

[0195] AGCG

[0196] AGCG

[0197] A probability model according to specific embodiments of the invention must decide among a number of competing hypotheses about the true template. For example, in attempting to decide between a T and an A at the highlighted position the model asks which event is more likely, that a T base generates an emission that is called as an A, or that an A

base is called as a T. While a standard alignment approach of choosing the template that maximizes the likelihood still applies, in the present invention the $b_{ij}(O)$ that models the probability of an observation is a function not solely of the base identity, but is also extended to return a measure of the probability of observing a pulse and various of its associated features on that transition. In this example, if the T pulse has stored or associated with it observed features indicating it was a higher intensity, longer pulse (and therefore less likely to be misclassified), while the A pulse was weaker and briefer, these features would be included in the probability model with other alignment probabilities to determine whether T or A was more probable. Other features being equal, the probability of having misclassified the bright T pulse being generated from an A template location would be much smaller than the probability of the weak brief A pulse being generated from a T base, therefore the model would call T as the consensus in that position. Because the data analyzed during the consensus alignment phase includes a number of different physical parameters of identified pulses and overall reaction parameters, rather than just a single quality score, many different characteristics of a real-time incorporation sequencing reaction can be used in the predictive model. The predictive model can thus be trained to account for the probability that a detected pulse was due to a branch or a stick of a labeled nucleotide analog, probabilities of which will vary for different bases, as well as account for overall reaction quality features such as overall noise detected at a reaction location or overall confidence of spectral classifications at a reaction location.

[0198] In a particular example embodiment, each state of an example HMM models a location along the template DNA strand where the synthesizing polymerase will reside between incorporation events. Two classes of transitions that can occur from this state are (1) a "move" transition where the polymerase incorporates a base and proceeds one position along the template, with a probability denoted by $a_{i,i+1}$, and (2) a "stay" transition where the polymerase binds a nucleotide, but unbinds before the incorporation event (a "branch") or a labeled nucleotide "sticks" transiently to the surface of the ZMW, inside the illumination region, and the polymerase does not move along the template, with probability given by $a_{i,i}$. A branch generally emits the symbol corresponding to the current template location while a stick generates a random symbol. The probability of branching and sticking are modeled as a function of the observation symbols (A C T G and null), and optionally modeled as a function of symbols for pulse metrics, such as intensity, duration, forward interval, subsequent interval, etc.

[0199] There are a variety of potential methods for generating the $b_{ij}(O)$ probability distribution for a multi-dimensional space of pulse parameters. Given the various pulse parameters and reaction parameters that may be calculated and stored as described herein, one presently preferred approach is to learn the distribution from empirical data acquired from known templates. By aligning the acquired pulse stream to the known template, pulses from a variety of classes can be used to generate empirical parameter distributions.

[0200] One method of scoring such a model during training is determining parameters that result in a maximum alignment length as is understood in the art.

Embodiment in a Programmed Information Appliance

[0201] FIG. 40 is a block diagram showing a representative example logic device in which various aspects of the present

invention may be embodied. As will be understood to practitioners in the art from the teachings provided herein, the invention can be implemented in hardware and/or software. In some embodiments of the invention, different aspects of the invention can be implemented in either client-side logic or server-side logic. As will be understood in the art, the invention or components thereof may be embodied in a fixed media program component containing logic instructions and/or data that when loaded into an appropriately configured computing device cause that device to perform according to the invention. As will be understood in the art, a fixed media containing logic instructions may be delivered to a viewer on a fixed media for physically loading into a viewer's computer or a fixed media containing logic instructions may reside on a remote server that a viewer accesses through a communication medium in order to download a program component.

[0202] FIG. 40 shows an information appliance (or digital device) 700 that may be understood as a logical apparatus that can read instructions from media 717 and/or network port 719, which can optionally be connected to server 720 having fixed media 722. Apparatus 700 can thereafter use those instructions to direct server or client logic, as understood in the art, to embody aspects of the invention. One type of logical apparatus that may embody the invention is a computer system as illustrated in 700, containing CPU 707, optional input devices 709 and 711, disk drives 715 and optional monitor 705. Fixed media 717, or fixed media 722 over port 719, may be used to program such a system and may represent a disk-type optical or magnetic media, magnetic tape, solid state dynamic or static memory, etc. In specific embodiments, the invention may be embodied in whole or in part as software recorded on this fixed media. Communication port 719 may also be used to initially receive instructions that are used to program such a system and may represent any type of communication connection.

[0203] The invention also may be embodied in whole or in part within the circuitry of an application specific integrated circuit (ASIC) or a programmable logic device (PLD). In such a case, the invention may be embodied in a computer understandable descriptor language, which may be used to create an ASIC, or PLD that operates as herein described.

OTHER EMBODIMENTS

[0204] The invention has now been described with reference to specific embodiments. Other embodiments will be apparent to those of skill in the art. In particular, a viewer digital information appliance has generally been illustrated as a personal computer. However, the digital computing device is meant to be any information appliance for interacting with a remote data application, and could include such devices as a digitally enabled television, cell phone, personal digital assistant, etc.

[0205] Although the present invention has been described in terms of various specific embodiments, it is not intended that the invention be limited to these embodiments. Modification within the spirit of the invention will be apparent to those skilled in the art. In addition, various different actions can be used to effect a request for sequence data.

[0206] It is understood that the examples and embodiments described herein are for illustrative purposes and that various modifications or changes in light thereof will be suggested by the teachings herein to persons skilled in the art and are to be included within the spirit and purview of this application and scope of the claims.

[0207] All publications, patents, and patent applications cited herein or filed with this application, including any references filed as part of an Information Disclosure Statement, are incorporated by reference in their entirety.

What is claimed is:

1. A method of identifying a nucleotide in a nucleic acid sequence, comprising:

receiving an optical signal from a reaction that produces optical signals in response to incorporation of a nucleotide analog in a template dependent primer extension reaction;

identifying the optical signal as an incorporation signal;

identifying a type of nucleotide associated with the optical signal based upon a spectrally distinguishable signal profile of the optical signal;

identifying a base in the nucleic acid sequence from the identification of the nucleotide associated with an optical signal that is also an incorporation signal.

2. The method of claim 1, wherein the step of identifying the optical signal as an incorporation signal comprises correlating the signal with a plurality of pulse metrics indicative of an incorporation signal.

3. The method of claim 1, wherein at least one of the plurality of pulse metrics is selected from the group consisting of pulse width, pulse intensity, pulse area, signal to noise ratio, pulse to diffusion background ratio, spectral correlation coefficient to identified dye, spectrum centroid, pulse spacing, pulse shape, and pulse polarization.

4. The method of claim 1, wherein at least one of the plurality of pulse metrics is selected from pulse width, pulse height and pulse shape.

5. The method of claim 1, wherein the step of identifying the optical signal as an incorporation signal comprises identifying the signal that exceeds one or more of a signal intensity threshold and a signal duration threshold.

6. The method of claim 1, wherein the step of identifying the optical signal as an incorporation signal comprises identifying the signal that exceeds both of a signal intensity threshold and a signal duration threshold.

7. The method of claim 1, wherein the step of receiving the optical signal comprises translating a received optical signal into optical signal data, and wherein the optical signal data comprises a first component indicative of an intensity of the optical signal, and a second component indicative of a location in a detection system at which the optical signal was received.

8. The method of claim 1, wherein the receiving step comprises imaging the optical signal on pixels of an imaging detector, and wherein the first component of the optical signal data comprises signal intensity data from a first plurality of pixels, and the second component of the optical signal data comprises a position of the plurality of pixels in the imaging detector.

9. The method of claim 1, wherein the step of identifying the type of nucleotide associated with the optical signal comprises comparing the second component of the optical signal data to an optical signal data component indicative of a location in a detection system at which the optical signal was received, from a standard fluorescent label associated with the type of nucleotide.

10. The method of claim 1, further comprising the step of providing a set of first and second standard signal data components for each of a plurality of different standard fluorescent labels or fluorescently labeled nucleotides, wherein the

step of identifying the type of nucleotide associated with the optical signal from the reaction, comprises comparing at least one of the first and second components of the optical signal data to the set of first and second standard signal data components.

11. A method of identifying a nucleotide in a nucleic acid sequence, comprising:

detecting on at least a first location of an optical signal detector, an optical signal from a reaction that produces optical signals in response to incorporation of a nucleotide analog in a template dependent primer extension reaction, wherein the optical signal detector translates optical signals to optical signal data;

identifying whether the optical signal data represents an incorporation signal;

identifying a type of nucleotide associated with the optical signal based upon the location of the detector upon which the optical signal data was derived; and

identifying a base in the nucleic acid sequence from the identification of the nucleotide associated with an optical signal that is also an incorporation signal.

12. A computer implemented process for calling bases in a nucleic acid sequencing process, comprising:

receiving optical signal data detected from a reaction of a polymerase/template/primer sequence complex and a detectably labeled nucleotide;

identifying the optical signal data as corresponding to an incorporation event, if the optical signal data exceeds one or more of a signal intensity threshold and a signal duration threshold;

correlating the optical signal data to incorporation of a first nucleotide, based upon comparison of the optical signal data to a standard set of optical signal data from different detectable labels, wherein each type of nucleotide has a different detectable label associated therewith;

calling the base in the template sequence as complementary to the first nucleotide where the optical signal data corresponds to an incorporation event.

13. The process of claim 12, further comprising repeating the receiving, identifying, correlating and calling steps to call a plurality of contiguous bases in the template sequence.

14. A system for determining a sequence of nucleotides in a nucleic acid sequence, comprising:

a reaction vessel comprising reaction mixture disposed therein, the reaction mixture comprising a nucleic acid polymerase complexed with a template nucleic acid sequence and a primer sequence, and one or more labeled nucleotides, wherein the reaction mixture produces an optical signal when the nucleic acid polymerase reacts with the one or more labeled nucleotides to incorporate the nucleotide into an extended primer;

an optical train positioned to receive the optical signal from the reaction mixture and image the optical signal onto an imaging detector which translates the optical signal to optical signal data;

a computer operably coupled to the imaging detector to receive the optical signal data from the detector, and programmed to:

identify the optical signal data as corresponding to an incorporation event, if the optical signal data exceeds one or more of a signal intensity threshold and a signal duration threshold;

correlate the optical signal data to incorporation of a first nucleotide, based upon comparison of the optical signal

data to a standard set of optical signal data from different detectable labels, wherein each type of nucleotide has a different detectable label associated therewith; and call the base in the template sequence as complementary to the first nucleotide where the optical signal data corresponds to an incorporation event.

15. A method of identifying a base in a target nucleic acid sequence, comprising:

receiving a first optical signal pulse from a reaction that produces optical signal pulses in response to incorporation of a nucleotide analog in a template dependent primer extension reaction;

comparing the first optical signal pulse to a plurality of pulse metrics derived from signal pulses associated with incorporation of one or more different nucleotide analogs in template dependent primer extension reactions; and

identifying the first pulse as a pulse associated with incorporation of specific nucleotide analog.

16. The method of claim 15, wherein the plurality of pulse metrics comprises at least two different pulse metrics.

17. The method of claim 15, wherein the plurality of pulse metrics comprises at least 5 different pulse metrics.

18. The method of claim 15, wherein the plurality of pulse metrics comprises at least 10 different pulse metrics.

19. A method of identifying a nucleotide in a nucleic acid sequence using a computer system comprising:

at an optical detector pixel array, capturing optical signals from a plurality of locations of sequencing reactions that produce optical signals at incorporation of a labeled nucleotide analog in a template dependent primer extension reaction;

transmitting captured data to a computer system;

identifying pixel areas of said optical detector array that correspond to one or more individual reaction locations; determining one or more calibration parameters applicable to a plurality of reaction locations;

determining one or more individualized calibration parameters for one or more individual reaction locations;

for individual reaction locations, analyzing optical signals captured by corresponding pixels of said optical detector to identify a series of significant pulses;

for individual reaction locations, classifying pulses in said series of pulses as associated with a type of nucleotide based upon a spectrally distinguishable signal profile of the optical signal pulse;

for individual reaction locations, calling bases in the nucleic acid sequence from the identification of the nucleotide associated with an optical signal that is also an incorporation signal;

determining one or more consensus sequences of bases by analyzing called bases and identified pulses from a plurality of locations;

outputting one or more of pulse characteristics, called bases, or consensus sequences to another computer system or to a user.

20. The method of claim 19, wherein the step of capturing the optical signals comprises translating a received optical signal into optical signal data, and wherein the optical signal data comprises a first component indicative of an intensity of the optical signal, and a second component indicative of a location in a detection system at which the optical signal was received.

21. The method of claim 19, wherein the step of capturing the optical signals comprises capturing a sequence of digital image frames, wherein each frame comprises a plurality of pixel intensity and position values for multiple reaction locations.

22. The method of claim 19, wherein the step of determining one or more calibration parameters applicable to a plurality of reaction locations comprises one or more of:

determining an alignment by comparing a line through central image areas of a plurality of reaction locations to the alignment of pixels of said pixel detector;

determining one or more high resolution spectral templates by averaging spectral calibration data detected at a plurality of pixel areas;

determining one or more high resolution spectral templates by averaging spectral calibration data detected at a plurality of pixel areas during a sub-pixel tilt between said detector and said reaction locations;

determining one or more noise parameters;

determining one or more drift parameters.

23. The method of claim 19, wherein the step of determining one or more individualized calibration parameters for one or more individual reaction locations comprises one or more of:

determining individualized reference centroids for reaction locations;

determining individualized sub-pixel reference centroids for reaction locations;

determining individualized spectral templates for reaction locations;

determining individualized background levels for reaction locations;

determining individualized noise characteristics for reaction locations.

24. The method of claim 19, wherein the step of for individual reaction locations, analyzing optical signals captured by corresponding pixels of said optical detector to identify a series of significant pulses comprises one or more of:

using individualized spectral templates for individual reaction locations, extracting one or more spectral traces;

identifying significant pulses in said one or more spectral traces by analyzing at least two metrics of said traces;

determining start and stop times for pulses from said one or more spectral traces;

combining start and stop times for one or more spectral trace to identify significant pulses of captured optical signals;

using as two of said at least two metrics, trace amplitude above a base line and duration of amplitudes above a base line;

using as two of said at least two metrics, a signal-to-noise measure of trace above a threshold and duration of measure above a threshold;

using as two of said at least two metrics, a signal to diffusion background ratio measure of trace above a threshold and duration of measure above a threshold;

combining start and stop times for one or more spectral trace to identify background intervals of captured optical signals.

25. The method of claim 24, further comprising additional pulse recognition steps comprising one or more of:

determining a second set of spectral traces to detect spectrally merged pulses;

adjusting a base-line threshold and a duration threshold to detect lower amplitude longer duration pulses;
analyzing adjacent pulses to merge algorithmically branched pulses.

26. The method of claim 19, wherein the step of classifying pulses in said series of pulses as associated with a type of nucleotide based upon a spectrally distinguishable signal profile of the optical signal pulse further comprises:

temporally averaging pixels of an optical signal pulse to produce a static spectrally spread image of said pulse;
comparing said static spectrally spread image to a plurality of spectral calibration templates;
determining a best fit of said static spectrally spread image to said spectral calibration templates and classifying said pulse as the best-fit calibration template.

27. The method of claim 26, further wherein:
said plurality of spectral calibration templates are individualized spectral calibration templates for particular locations.

28. The method of claim 27, further wherein:
said plurality of individualized spectral calibration templates for a particular location is the same template used to extract traces for pulse identification.

29. The method of claim 19, wherein the step of capturing optical signals further comprises:

capturing a time series of static digital image frames of multiple reaction locations during incorporation reactions, said frames each comprising a first component indicative of an intensity of the optical signal, and a second component indicative of a location in a detection system at which the optical signal was received.

30. The method of claim 19, wherein the step of identifying pixel locations comprises:

for individual reaction locations, determining a spectral axis of a set of pixel locations corresponding to an individual reaction location, said spectral axis parallel to a spectral refraction of optical signals produced in response to incorporation of a nucleotide analog; and
determining a spatial axis of said set, said spatial axis orthogonal to said spectral axis.

31. The method of claim 30, further comprising:

for individual reaction locations, combining pixels in said spatial axis direction to produce a spectral images of said reaction locations, said spectral images comprising a plurality of pixels parallel to said spectral axis and a reduce number of one or more pixels parallel to said spatial axis.

32. The method of claim 31 wherein said combining comprises:

at said optical detector, storing parameters indicating pixel locations of said optical detector array that correspond to one or more individual reaction locations;
at said optical detector, storing parameters indicating how pixels should be combined;
combining pixels prior to said transmitting.

33. The method of claim 31 wherein said combining comprises:

at said computer system, combining pixels in said spatial direction using weighted values to maximize detected flux.

34. The method of claim 24, further wherein:

said extracting one or more spectral traces comprises at each frame, analyzing said plurality of pixel levels to determine one or more spectral intensity levels.

35. The method of claim 34, further wherein:

said one or more spectral intensity levels comprises one to four spectral intensity levels.

36. The method of claim 24, wherein extracting one or more spectral traces from said at least one optical signal comprises comparing a position component of the optical signal data to an optical signal data component indicative of a location in a detection system at which the optical signal was received, from a standard fluorescent label associated with the type of nucleotide.

37. The method of claim 19, further wherein said calling-bases comprises:

identifying incorporation events from one or more of said significant pulses of said optical signals.

38. The method of claim 37, wherein the step of identifying incorporation events from optical signals comprises correlating the optical signals with a plurality of pulse metrics indicative of incorporation events.

39. The method of claim 38, wherein at least one of the plurality of pulse metrics is selected from the group consisting of pulse width, pulse intensity, pulse area, signal to noise ratio, pulse to diffusion background ratio, spectral correlation coefficient to identified dye, spectrum centroid, pulse spacing, pulse shape, and pulse polarization.

40. The method of claim 38, wherein at least one of the plurality of pulse metrics is selected from pulse width, pulse height and pulse shape.

41. The method of claim 37, wherein the step of identifying incorporation events from optical signals comprises identifying the signal that exceeds one or more of a signal intensity threshold and a signal duration threshold.

42. The method of claim 37, wherein the step of identifying incorporation events from optical signals comprises identifying the signal that exceeds both of a signal intensity threshold and a signal duration threshold.

43. The method of claim 38, wherein the plurality of pulse metrics comprises at least two different pulse metrics.

44. The method of claim 38, wherein the plurality of pulse metrics comprises at least 5 different pulse metrics.

45. The method of claim 38, wherein the plurality of pulse metrics comprises at least 10 different pulse metrics.

46. A method of identifying a nucleotide in a nucleic acid sequence using a computer system comprising:

capture pixels from a pixel detector for multiple reaction locations;

determine correlations (gridding) between pixel sets and reaction locations;

perform one or more individualized calibrations and store one or more individualized reaction location calibration parameters and/or data sets

perform one or more collective array calibrations and store one or more collective calibration parameters;

determine one or more data reduction parameters;

capture a time sequence of pixel images from a pixel detector for multiple reaction locations of a sequencing array during multiple sequencing reactions;

perform one or more data reductions on captured pixels; for individual reaction locations, extract dye separated spectral traces;

analyze extracted dye separated traces to determine trace pulse characteristics;

use trace pulse start and end times (optionally from multiple traces) to identify a pulse in data reduced captured pixel data compare a pulse in data reduced captured pixel data to one or more spectral templates to classify captured pulses as associated with a particular dye-label; perform an initial base-calling for individual reaction locations;

examine pulse data from multiple locations with overlapping sequence data and perform one or more sequence

alignments using one or more of initial base calling from individual reaction locations, and probabilities associated with one or more of a classification, a false-positive pulse detection, a false-negative pulse detection; and output one or more sequences produced by an alignment of pulse data, classification data, and probability data from multiple reaction locations, where output can include presentation to a user on a display or printout, recording sequence data on a tangible medium, or transmitting sequence data to an external computer system.

* * * * *