

Robert Lyons, Ph.D., Director

***Quality Control and
Crowd Control in the
DNA Sequencing Core***

This talk is intended to show methods by which the University of Michigan DNA Sequencing Core maintains high quality of its data product while increasing its capacity to meet the immense demand.

The Core needs to:

- Simplify data handling, minimize the human time expended on quality assessment.
- Respond to client concerns, interact with them and talk about *their* data, *their* problems.
- Notice and troubleshoot problems quickly, even *before* the client does.

With the ongoing pressure to reduce costs, it behooves a Core to expand. The primary benefit: improved cost efficiency. The drawback: difficulty in monitoring and maintaining the quality of data you produce. With growth, it also becomes increasingly difficult to respond to client concerns, especially when they feel you have erred.

The Core needs to:

- Simplify data handling, minimize the human time expended on quality assessment.
- Respond to client concerns, interact with them and talk about *their data, their problems*.
- Notice and troubleshoot problems quickly, even *before* the client does.

First we will discuss some scenarios in which you need to respond to client concerns regarding your data quality.

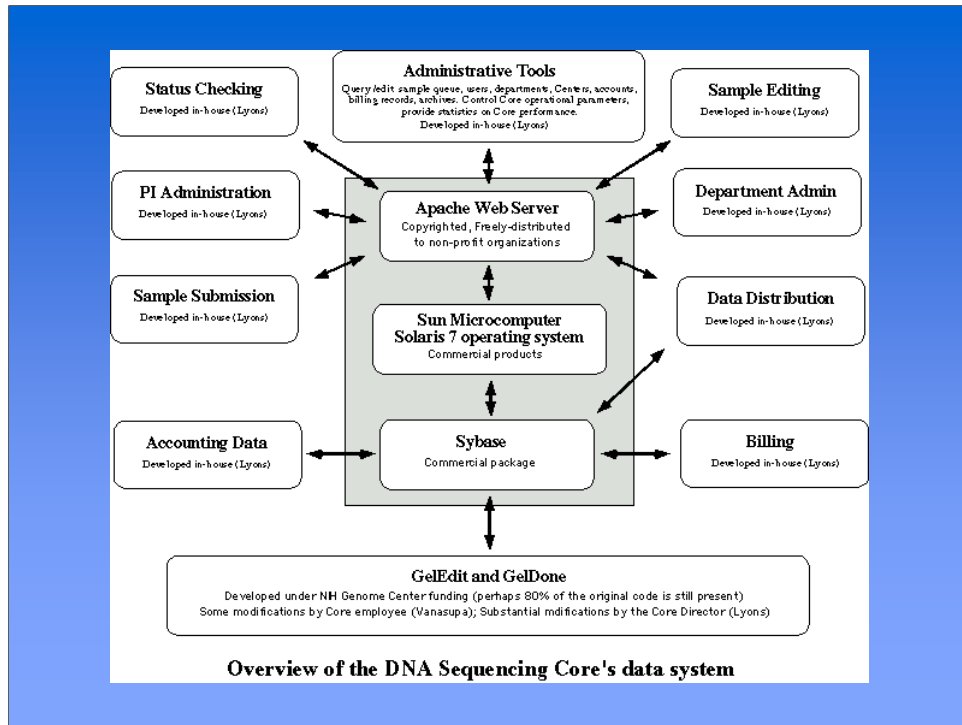


“Why did this *one* sample fail, when all the rest worked?”

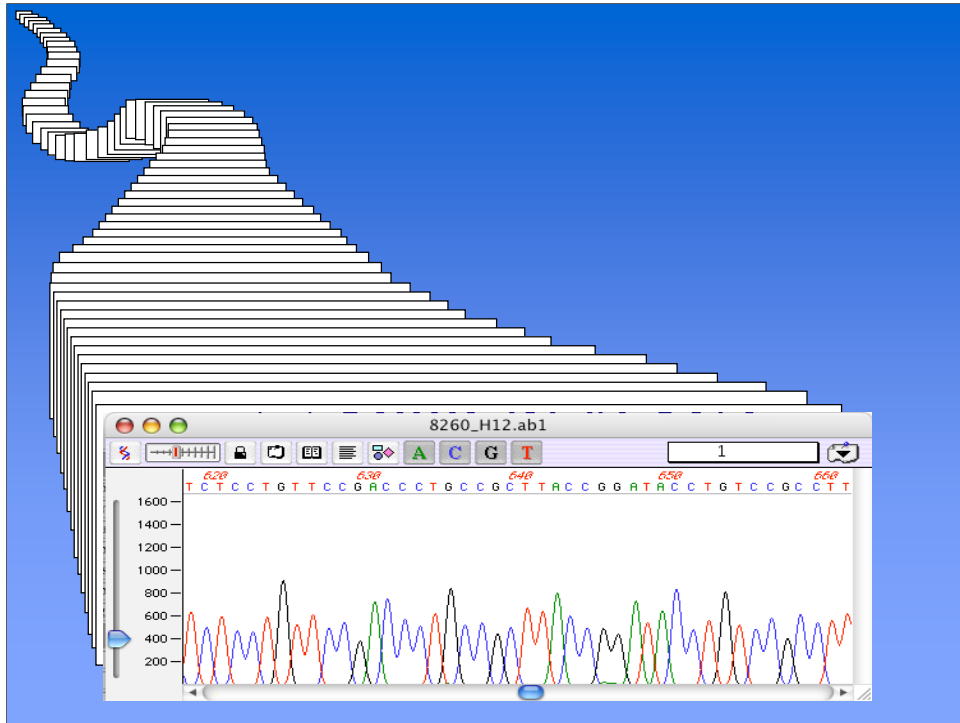
“You guys can’t sequence mini-preps!”

“You guys can’t handle custom primers!”

A very typical scenario: a client wonders why one sample failed when all the rest seemed (to him) to have worked. How can you address him most effectively (and quickly, too)?



This talk will show off numerous features of our Laboratory Information Management System. We will not be discussing technician features of that system; only its operational advantages. If you want more detail, please see: <http://seqcore.brcf.med.umich.edu/doc/dnaseq/datasys.html>



The Gold Standard method for assessing what happened to a client's samples is to examine lots and lots of chromatograms. You need a way to quickly call up any chromatogram on demand ... the faster, the better. Our system has some half million chromatograms that can be retrieved and displayed within seconds of recognizing the need to do so. **HOWEVER**, we are discussing how to address a client's concerns as quickly as possible, and in this case, there is a much quicker way.

List of Lanes: Click on the lane number to see complete details on that lane.

Lane:	Well:	Sampleno:	Primer:	PILogin:	Login:	ServiceDate:	Status:	Comments:	phred20:	AvgPht:	CapNo:	Signal:
1	A01	pGEM3	T7	Standard					931	56	15	747
2	A02	347665	M13FORW	hildebrand	otto	Jan 20, 2004	d		1122	47	16	149
3	A03	347665	M13REV	hildebrand	otto	Jan 20, 2004	d	Enzyme slipped @ base 100 after homo-polymer region.	827	49	31	115
4	A04	347665	T7	hildebrand	otto	Jan 20, 2004	d		900	58	32	165
5	A05	347687	T7	pichersky	ijijima	Jan 20, 2004	d		947	54	47	112
6	A06	347687	custom1	pichersky	ijijima	Jan 20, 2004	d		841	59	48	111
7	A07	347688	T7	pichersky	ijijima	Jan 20, 2004	d		892	57	63	120
8	A08	347689	T7	pichersky	ijijima	Jan 20, 2004	d		896	55	64	100
9	A09	347694	custom1	robins	krebs	Jan 20, 2004	d		899	57	79	230
10	A10	347695	custom1	robins	krebs	Jan 20, 2004	d		946	55	80	208
11	A11	347696	custom1	robins	krebs	Jan 20, 2004	d		886	57	95	486
12	A12	347697	custom1	robins	krebs	Jan 20, 2004	d	Elevated background.	814	52	96	58
13	B01	347647	custom1	hildebrand	utsch	Jan 20, 2004	d		805	59	13	1032
14	B02	347648	custom1	hildebrand	utsch	Jan 20, 2004	d	Superimposed sequences @ start.	770	60	14	952
15	B03	347649	custom1	hildebrand	utsch	Jan 20, 2004	d		804	60	29	809
16	B04	347650	custom1	hildebrand	utsch	Jan 20, 2004	d		798	60	30	737
17	B05	347651	custom1	hildebrand	utsch	Jan 20, 2004	d		805	59	45	523
18	B06	347652	custom1	hildebrand	utsch	Jan 20, 2004	d		803	59	46	553
19	B07	347653	custom1	hildebrand	utsch	Jan 20, 2004	d		797	59	61	443
20	B08	347654	custom1	hildebrand	utsch	Jan 20, 2004	d		805	59	62	494
21	B09	347655	custom1	hildebrand	utsch	Jan 20, 2004	d	Elevated background @ start of sequence.	788	60	77	619
22	B10	347656	custom1	hildebrand	utsch	Jan 20, 2004	d		804	57	78	313
23	B11	347741	M13REV	clark	prigge	Jan 20, 2004	d		862	58	93	98
24	B12	347741	T7	clark	prigge	Jan 20, 2004	d		886	55	94	150

After we process a set of 96 samples, we can view a screen of information like this. It includes sample/client tracking information (on the left), technician comments for each lane (center) and statistical assessment of each lane's outcome.

edate:	Status:	Comments:	phred20:	AvgPhd:	CapNo:	Signal:
			931	56	15	747
, 2004	d		1122	47	16	149
, 2004	d	Enzyme slipped @ base 100 after homo-polymer region.	827	49	31	115
, 2004	d		900	58	32	165
, 2004	d		947	54	47	112
, 2004	d		841	59	48	111
, 2004	d		892	57	63	120
, 2004	d		896	55	64	100
, 2004	d		899	57	79	230

Zooming on on the statistical columns: 'phred20' is the read length (derived from phred basecaller via our program 'fillstats'), followed by 'AvgPhd', the average phred Q score over that read length. The last column is the 'G' signal for that lane. Now we'll see how to use these statistics to quickly dismiss our client's suggestion that we screwed up...

49	E01	332024	custom1	reisman	johnson	Nov 11, 2003	d		430	60	7	184
50	E02	332024	custom2	reisman	johnson	Nov 11, 2003	d		422	61	8	589
51	E03	332025	custom1	reisman	johnson	Nov 11, 2003	d		432	60	23	681
52	E04	332025	custom2	reisman	johnson	Nov 11, 2003	d		422	62	24	821
53	E05	332033	M13FORW	xxxxxxxxxx	xxxxx	Nov 11, 2003	d	Drop in signal at approx. base 30. NOT RELIABLE DATA thereaf	251	20	39	24
54	E06	332033	M13REV	xxxxxxxxxx	xxxxx	Nov 11, 2003	d	Drop in signal at approx. base 60. NOT RELIABLE DATA thereaf	222	25	40	26
55	E07	332034	M13FORW	xxxxxxxxxx	xxxxx	Nov 11, 2003	d	Drop in signal at approx. base 30. NOT RELIABLE DATA thereaf	8	12	55	38
56	E08	332034	M13REV	xxxxxxxxxx	xxxxx	Nov 11, 2003	d	Drop in signal at approx. base 60. NOT RELIABLE DATA thereaf	312	36	56	37
57	E09	332035	M13FORW	xxxxxxxxxx	xxxxx	Nov 11, 2003	d	Drop in signal at approx. base 30. NOT RELIABLE DATA thereaf	313	20	71	26
58	E10	332035	M13REV	xxxxxxxxxx	xxxxx	Nov 11, 2003	d	Drop in signal at approx. base 60. NOT RELIABLE DATA thereaf	317	33	72	33
59	E11	332036	M13FORW	xxxxxxxxxx	xxxxx	Nov 11, 2003	d	Drop in signal at approx. base 30. NOT RELIABLE DATA thereaf	11	9	87	25
60	E12	332036	M13REV	xxxxxxxxxx	xxxxx	Nov 11, 2003	d	Drop in signal at approx. base 60. NOT RELIABLE DATA thereaf	358	41	88	41
61	F01	332037	M13FORW	xxxxxxxxxx	xxxxx	Nov 11, 2003	d	No bands, no data	135	17	5	??
62	F02	332037	M13REV	xxxxxxxxxx	xxxxx	Nov 11, 2003	d	Drop in signal at approx. base 60. NOT RELIABLE DATA thereaf	14	11	6	23
63	F03	332038	M13FORW	xxxxxxxxxx	xxxxx	Nov 11, 2003	d	Drop in signal at approx. base 30. NOT RELIABLE DATA thereaf	0	10	21	22
64	F04	332038	M13REV	xxxxxxxxxx	xxxxx	Nov 11, 2003	d	Drop in signal at approx. base 60. NOT RELIABLE DATA thereaf	261	31	22	33
65	F05	332039	M13FORW	xxxxxxxxxx	xxxxx	Nov 11, 2003	d	Drop in signal at approx. base 30. NOT RELIABLE DATA thereaf	382	34	37	30
66	F06	332039	M13REV	xxxxxxxxxx	xxxxx	Nov 11, 2003	d	Drop in signal at approx. base 60. NOT RELIABLE DATA thereaf	371	34	38	34
67	F07	332040	M13FORW	xxxxxxxxxx	xxxxx	Nov 11, 2003	d	Drop in signal at approx. base 30. NOT RELIABLE DATA thereaf	367	30	53	31
68	F08	332040	M13REV	xxxxxxxxxx	xxxxx	Nov 11, 2003	d	Drop in signal at approx. base 60. NOT RELIABLE DATA thereaf	380	37	54	37
69	F09	332041	M13FORW	xxxxxxxxxx	xxxxx	Nov 11, 2003	d	Drop in signal at approx. base 30. NOT RELIABLE DATA thereaf	335	23	69	33
70	F10	332041	M13REV	xxxxxxxxxx	xxxxx	Nov 11, 2003	d	Drop in signal at approx. base 60. NOT RELIABLE DATA thereaf	322	33	70	39

Here's a display with our client's samples. X's in the client columns hide his true identity. Note first that the technician comments warned the client that the results were NOT RELIABLE. The reason: the 'G' signal on all of Xxxx's samples is far, far too low. In other words, this wasn't a set where all samples worked except for one In fact, all samples worked POORLY, and one worked slightly MORE poorly. Case dismissed.

The Core Director can get statistics on:

- Which technician generates the best data?
- Which sequencer is having problems?
- Do specific lanes have anomalous failure rates?
- Have we had a decline on quality since *X*?
- What vectors sequence poorly?

Using the phred-derived statistics stored in our database for each lane, we can quickly identify technicians who are producing poor data, sequencers that need attention or clients who are doing sub-standard work.

“Why did this *one* sample fail, when all the rest worked?”



“You guys can’t sequence mini-preps!”

“You guys can’t handle custom primers!”

Better yet, we can average out the read length statistics and compare different clients, revealing failure patterns that are quite useful. Here’s another example of a client who complained about the Core’s work.

Results for pilogin: jones										
Begin Date: January 01, 2003					End Date: January 01, 2005					
Researcher	Purified Plasmid		Large DNA		Mini Prep		PCR Product		Single Strand	
	Samples	Avg Phred	Samples	Avg Phred	Samples	Avg Phred	Samples	Avg Phred	Samples	Avg Phred
swift	16	846			30	626	55	569		
tolstoy				→	236	838				
clark	90	578					4	303		
craig					20	390				
barry				→	68	475				
adams	9	212			15	126				
miller	16	829								
palmer	9	185			20	426				
bender	27	656			26	817				
keller	115	725			43	591	71	554		
mitchell					191	519	14	620		
royce					46	538				
evans					61	831				

Although he claims his failed miniprep sequences were our fault, it was easy to prove otherwise. Here's a display of the average read length for several of his lab-mates (names have been changed). Note that our plaintiff "Mr. Barry" indeed has a poor average read length for mini-preps. However, clearly we can sequence minipreps quite well, since "Mr. Tolstoy", from the same lab, has an outstanding read length for minipreps!

“Why did this *one* sample fail, when all the rest worked?”

“You guys can’t sequence mini-preps!”



“You guys can’t handle custom primers!”

Another example: someone who claims we can’t sequence using client-provided (custom) primers. Looking at the statistics

Quality statistics from the query you posted:

pilgin	login	Sample Count	Avg phred20:	Avg 'avgphd':
hume	friday	221	877	56
hanna	heffernan	212	852	54
mapp	lum	245	846	52
hildebrand	wolf	462	838	51
guan	yoli	227	792	50
zyang	dyang	484	786	55
davis	davis	1586	781	49
saltiel	bao	259	772	48
nunez	nishito	490	765	50
	nalmbo			
swindell	maxwell	82	493	36
moore	fisher	84	493	33
peltier	erikson	102	445	31
shane	hiram	100	440	33
dirk	wyeth	62	424	30
esteban	jung	80	417	30
moyer	elders	104	402	26
mellen	stande	48	367	26
cole	shindler	93	268	29

The top of this table shows the real names of *competent* labs, those who have an excellent record when sequencing with custom primers. They are all highly-respected labs. Clearly we CAN sequence with custom primers. At the bottom of the list, however, is the problem client (names changed in this half of the list). Clearly the problem is only in HIS lab. Case dismissed.

The Core Director can get statistics on:

- Which technician generates the best data?
- Which sequencer is having problems?
- Do specific lanes have anomalous failure rates?
- Have we had a decline in quality since June?
- What vectors sequence poorly?

Using the phred-derived statistics stored in our database for each lane, we can quickly identify technicians who are producing poor data, sequencers that need attention or clients who are doing sub-standard work.

We save “flags” for each sample, recording:

- **Success vs Failure**
- **“Set-Mode” vs Isolated**

When ‘FillStats’ assesses the statistics for each lane, it also sets two flags: a ‘success’ flag indicates whether the lane read out as far as expected (it adjusts expectations for PCR products vs plasmids). A ‘Set-Mode’ flag has to be explained later ...

DNA Sequencing Core

Gel Number: 8222

Instrument: James

Technician: Cesar Rowley

Run Date: Feb 12, 2004

Gel done?: Yes

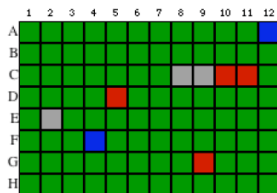
Lanes: 96

Date Done: Feb 16, 2004

Comments: BDT v1.1; 1x seq buffer POP-7 princeton columns.

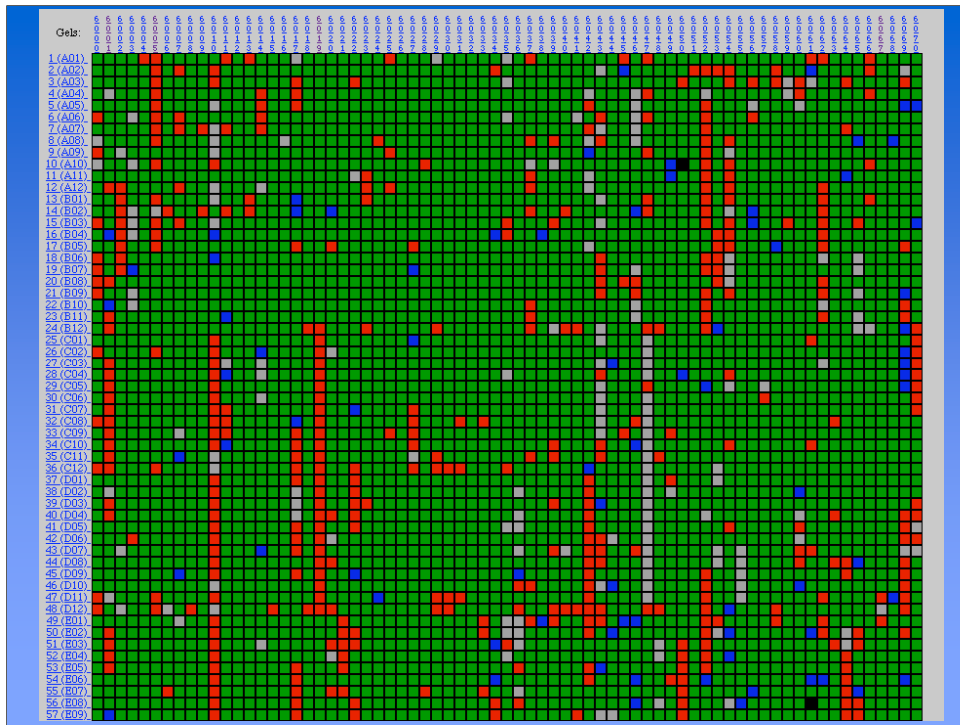
LEGEND:

■	Successful sequencing
■	Partial success
■	Failed sequencing (<100 bp)
■	Repeat suggested
■	No assessment available

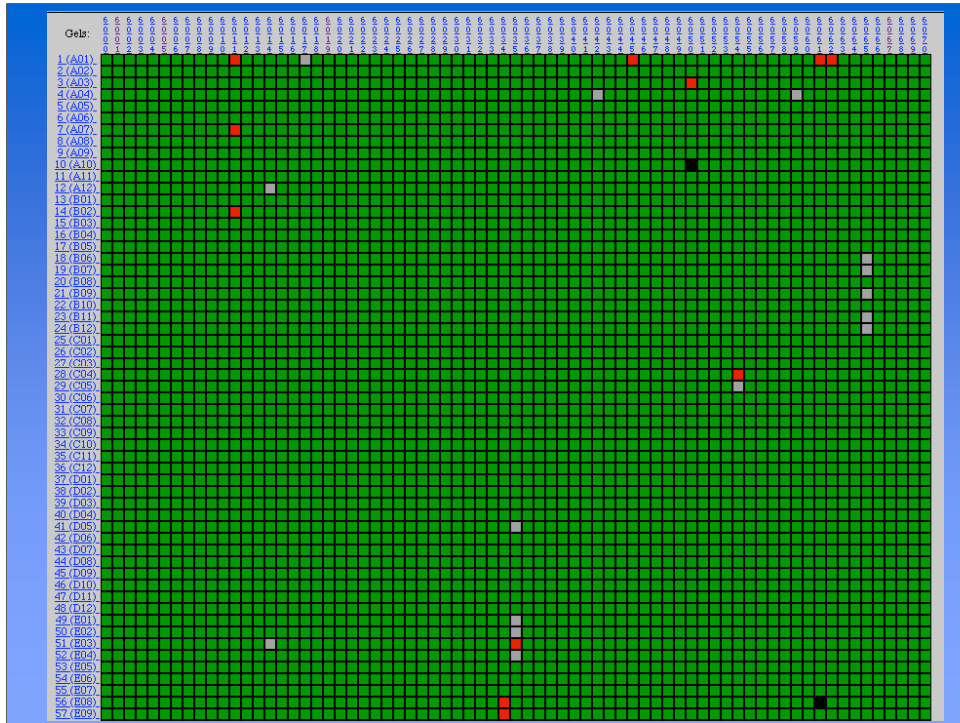


[Back to Admin Page](#)

With the 'Success' flag, we can display matrices of samples, color coded for success vs failure (or partial success). This display isn't all that useful, since we can comprehend the statistics on a 96-well sample set pretty easily.



With a larger set of samples, the numerical stats are impossible to grasp, so the success/fail flag becomes far more useful. Note the clear pattern of vertical red streaks. These are samples that all arrived together as a set, and (being identically prepared) all failed in the same way. This is common, and we call the phenomenon a “set-mode” failure. By identifying “set-mode” failures, we can do an interesting QC exercise...



Here, we've blanked out all the "set-mode" failures, plus all the Core-identified repeats and samples that failed for obvious reasons (Taq slip, sec-structure, etc). What's left are failures where we can't explain away the cause ... in other words possible Core screw-ups! In fact, I do this when I want to try to identify places where my technicians have made mistakes. By the way, in this case, none of these remaining colored cells were in fact screw-ups.

DNA Sequencing Core

Gel Number: 8222

Instrument: James

Technician: Cesar Rowley

Run Date: Feb 12, 2004

Gel done?: Yes

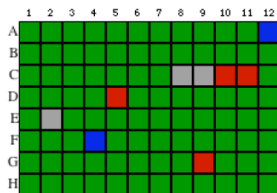
Lanes: 96

Date Done: Feb 16, 2004

Comments: BDT v1.1; 1x seq buffer POP-7 princeton columns.

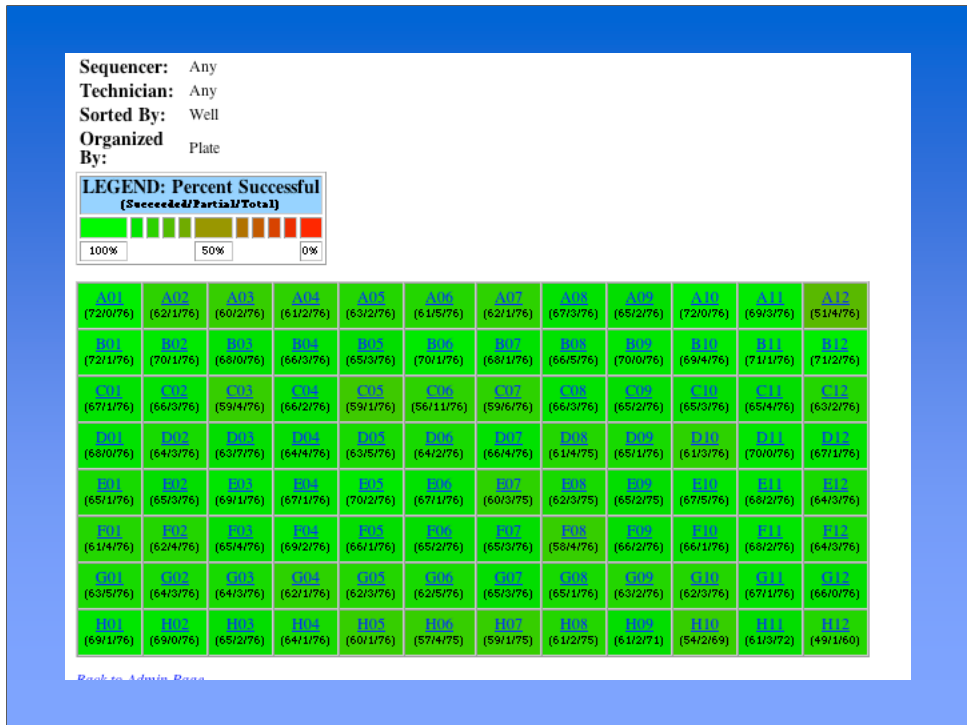
LEGEND:

■	Successful sequencing
■	Partial success
■	Failed sequencing (<100 bp)
■	Repeat suggested
■	No assessment available



[Back to Admin Page](#)

We can also display those success/fail flags in a different format - arranged as 96-well. Here's one I've shown previously ... again, it's not very useful.



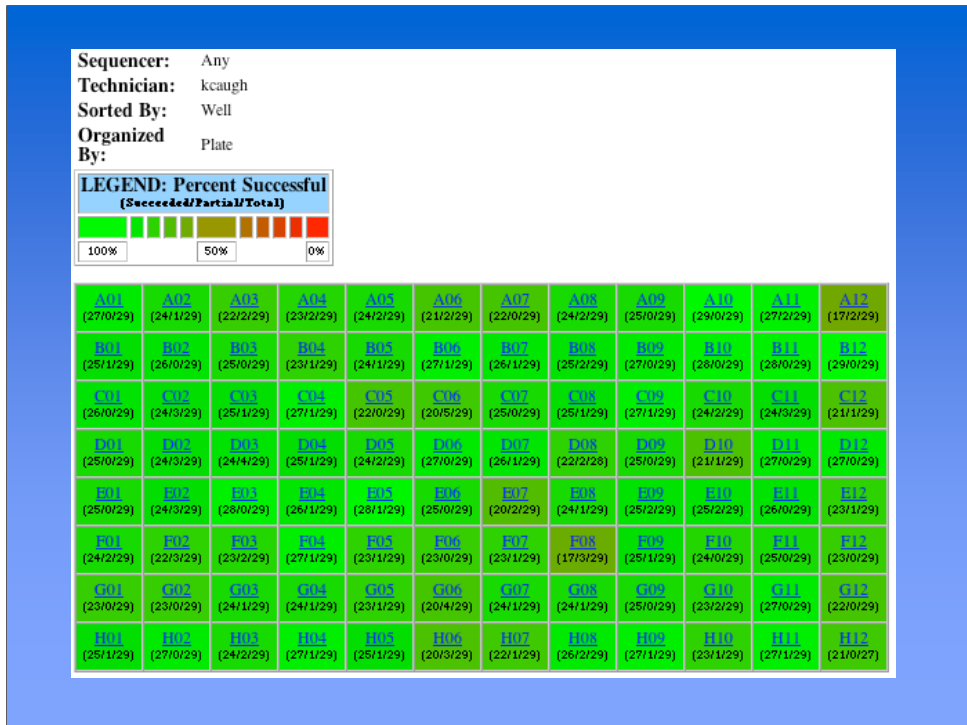
When we include two hundred plates, we can color-code each position of the 96-well grid to display the rate of failure at each position. Note the slightly more red tone of well A12. The capillary for that well is partially plugged, and samples there tend to fail more often than others. We use this ‘plate averaged’ display to spot patterns of failure.

Quality Control Displays		
Please enter the desired gel number(s) below: (To see just one gel, enter its number in the first blank)		
Starting Gelno:	<input type="text" value="8050"/>	Enter a single gel number here to view one gel, or enter the first of a series of gels you want to see.
Ending Gelno: <i>(optional)</i>	<input type="text" value="8200"/>	To view a range of gels, enter the last gel number here; otherwise, leave this blank (to see a single gel).
All gels or Selected ones? <i>(Ignored for single-gel requests)</i>	<input type="text" value="All Gels"/>	If you choose 'Selected Gels' you'll be able to go through a list clicking checkboxes for the gels you want to see.
Sort Display By:	<input type="text" value="Well"/>	Choosing 'Well' shows you the wells from A01 to H12. Choosing 'Caps' lists the capillary numbers (in order) instead.
Organize Display As: <i>(Single gels display as 'Plate')</i>	<input type="text" value="Plate (averaged results)"/>	You can see a Grid of colors showing success or failure for each lane, or you can get a display showing the percent success rate "averaged" over all gels, arranged in a 12 x 8 'Plate'-type view.
Limit to Sequencer: <i>(Ignored for single-gel requests)</i>	<input type="text" value="Any"/>	Select one sequencer to see quality control data for only that one machine.
Limit to Technician: <i>(Ignored for single-gel requests)</i>	<input type="text" value="Any"/>	Select a single technician to see quality control data for just that one individual.
Click to submit: <input type="button" value="Submit Query"/>		

We can select what “gels” (sample sets) to include in the display, we can limit it to specific sequencers or specific technicians. For example ...

Quality Control Displays		
Please enter the desired gel number(s) below: (To see just one gel, enter its number in the first blank)		
Starting Gelno:	<input type="text" value="8050"/>	Enter a single gel number here to view one gel, or enter the first of a series of gels you want to see.
Ending Gelno: <i>(optional)</i>	<input type="text" value="8200"/>	To view a range of gels, enter the last gel number here; otherwise, leave this blank (to see a single gel).
All gels or Selected ones? <i>(Ignored for single-gel requests)</i>	<input type="text" value="All Gels"/>	If you choose 'Selected Gels' you'll be able to go through a list clicking checkboxes for the gels you want to see.
Sort Display By:	<input type="text" value="Well"/>	Choosing 'Well' shows you the wells from A01 to H12. Choosing 'Caps' lists the capillary numbers (in order) instead.
Organize Display As: <i>(Single gels display as 'Plate')</i>	<input type="text" value="Plate (averaged results)"/>	You can see a Grid of colors showing success or failure for each lane, or you can get a display showing the percent success rate "averaged" over all gels, arranged in a 12 x 8 'Plate'-type view.
Limit to Sequencer: <i>(Ignored for single-gel requests)</i>	<input type="text" value="Any"/>	Select one sequencer to see quality control data for only that one machine.
Limit to Technician: <i>(Ignored for single-gel requests)</i>	<input type="text" value="Kayce Caugh"/>	Select a single technician to see quality control data for just that one individual.
Click to submit: <input type="button" value="Submit Query"/>		

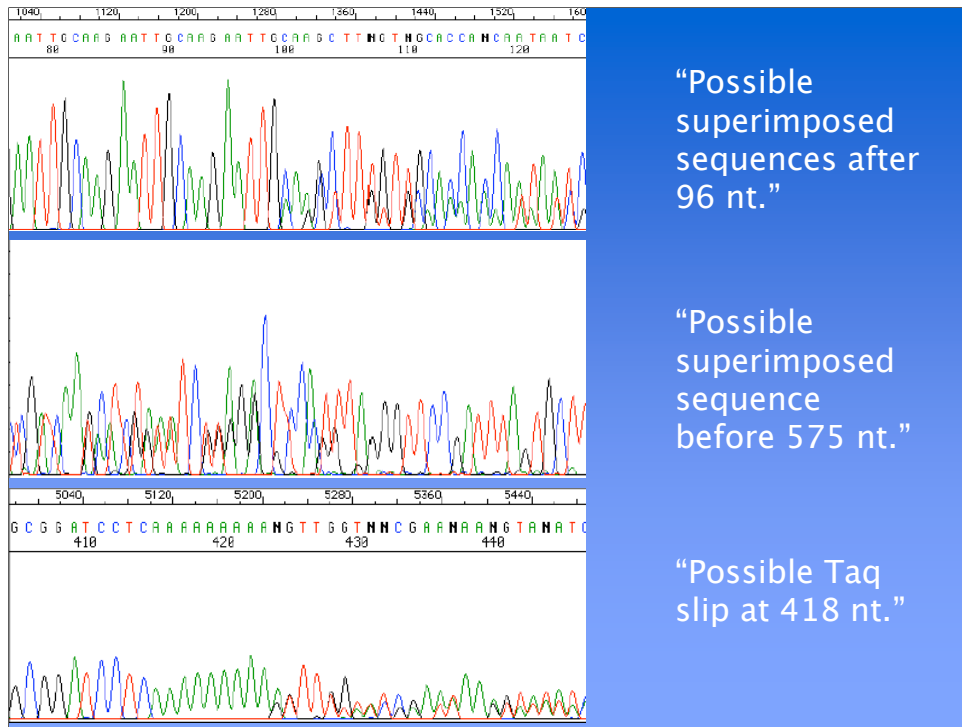
... here we are going to limit the results to show only sample sets that were processed by my technician, Kayce Caugh. Submitting this form gives ...



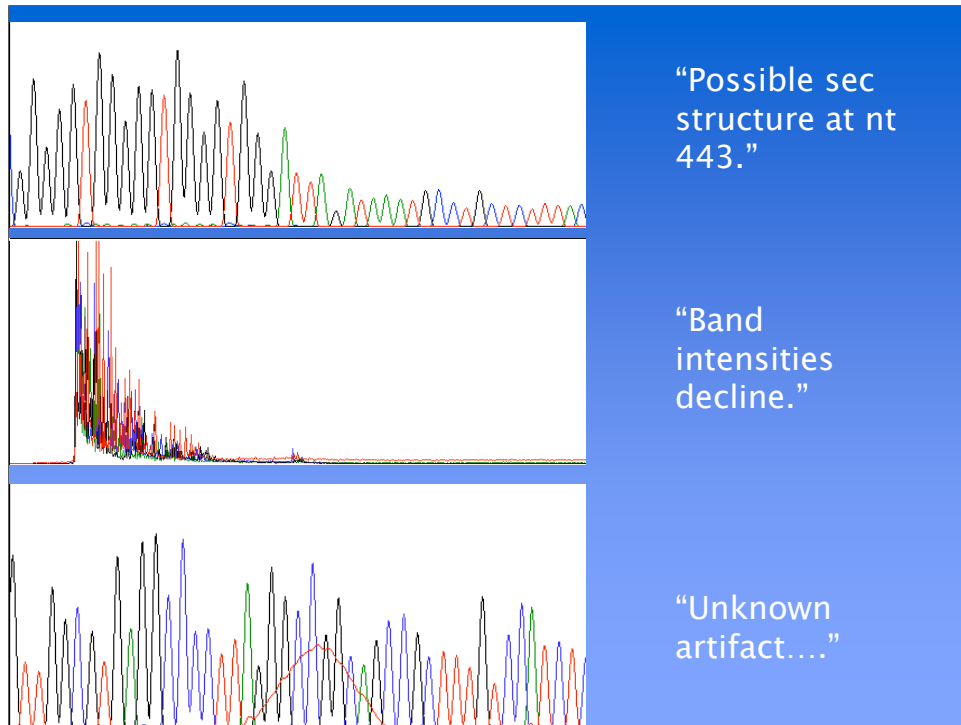
... this Q/C display. Note the A12 anomaly is still there. Note, however, that F08 is now distinctly red. Ms. Caugh was unaware of this point of failure, and had no explanation. When we blew compressed air into all the PCR heads she tends to use, this F08 anomaly went away.

Average Phred20 length, entire gel: 725											Average Phred20 for 'good' lanes: 781			
List of Lanes: Click on the lane number to see complete details on that lane.														
Lane	Well	Sampleno	Primer	PILogin	Login	Servicedate	Status	TechComments	AutoComments	phred20	AvgPhd	CapNo	Signal	
1	A01	pGEMEX	T3	Standard						936	57	15	502	
2	A02	354262	custom1	zhang	cho	Feb 12, 2004	d	No bands, no data.	No bands.	0	8	16	18	
3	A03	354263	custom1	zhang	cho	Feb 12, 2004	d	No bands, no data.	No bands.	0	8	31	17	
4	A04	354264	custom1	zhang	cho	Feb 12, 2004	d	No bands, no data.	No bands.	0	8	32	18	
5	A05	354265	custom1	zhang	cho	Feb 12, 2004	d	No bands, no data.	No bands.	0	7	47	16	
6	A06	354274	T3	hildebrand	raymond	Feb 12, 2004	d	Very faint/weak. Dye residue at the start. Very elevated bac	Very weak bands.	548	35	48	24	
7	A07	354274	T7	hildebrand	raymond	Feb 12, 2004	d	Very faint/weak. Dye residue at the start. Very elevated bac	Very weak bands.	749	41	63	31	
8	A08	354275	custom1	lieberman	pacheco	Feb 12, 2004	d			846	60	64	175	
9	A09	354275	custom2	lieberman	pacheco	Feb 12, 2004	d			882	60	79	201	
10	A10	354275	custom3	lieberman	pacheco	Feb 12, 2004	d			953	57	80	243	
11	A11	354275	custom4	lieberman	pacheco	Feb 12, 2004	d			831	61	95	694	
12	A12	354276	custom1	lieberman	pacheco	Feb 12, 2004	d	Elevated background: see chromatogram for base calling. We w	Somewhat weak bands. Unknown artifact from 373 to 389 nt.	790	50	96	86	
13	B01	354276	custom2	lieberman	pacheco	Feb 12, 2004	d			918	59	13	250	
14	B02	354276	custom3	lieberman	pacheco	Feb 12, 2004	d			952	57	14	326	
15	B03	354276	custom4	lieberman	pacheco	Feb 12, 2004	d			942	57	29	537	
16	B04	354277	T7	gantz	lai	Feb 12, 2004	d		Band intensities decline.	922	58	30	263	
17	B05	354277	custom1	gantz	lai	Feb 12, 2004	d	Slightly elevated background.		1037	49	45	158	
18	B06	354278	T7	gantz	lai	Feb 12, 2004	d		Band intensities decline.	911	58	46	191	
19	B07	354278	custom1	gantz	lai	Feb 12, 2004	d	Slightly elevated background.		844	58	61	187	

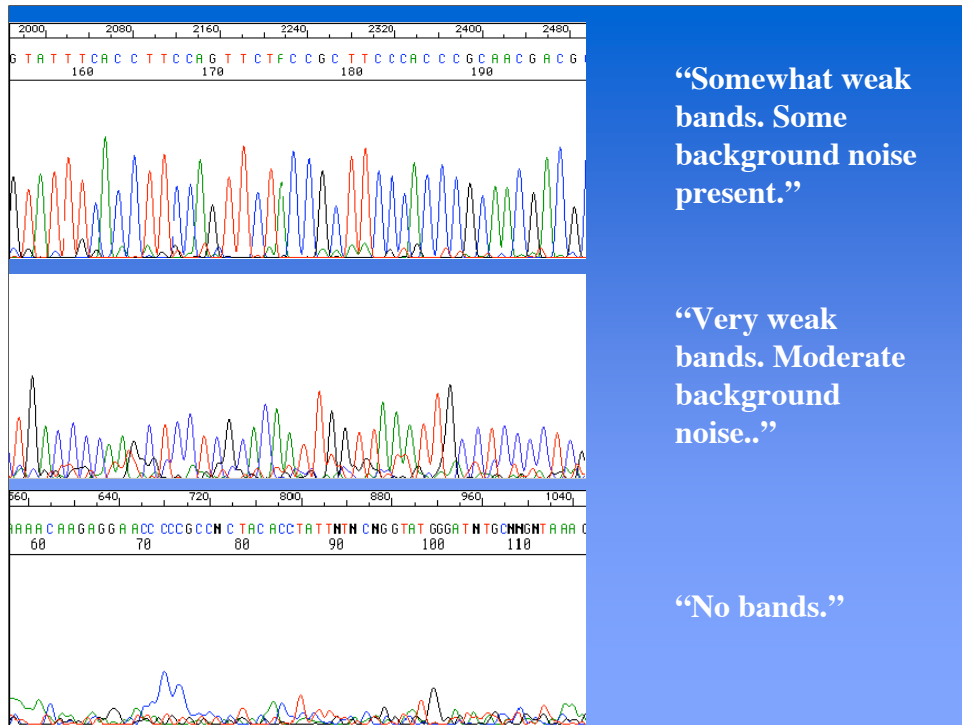
Change of subject: Previous LIMS features tended to benefit the Core Director more than anyone else. Here's a feature that benefits the technicians. Note the extra "Comment" field, called "AutoComments". These are computer generated by a program that scans the chromatograms lane by lane and attempts to diagnose the problems it finds.



The AutoComment program can identify quite a number of common failure modes for sequencing, including two clones mixed together (top), two PCR fragments mixed together (middle) and Taq slip (bottom).



The AutoComment program also knows what a secondary-structure artifact looks like, what the salt effect looks like, and it recognizes (but does not yet correctly diagnose) dye blobs.



AutoCommenting also can comment on the relative weakness/strength of lanes (including ‘Grossly overloaded’, not shown) and knows what a blank lane is.

Automatic Chromatogram Commenting:

- A supervisor can check the accuracy of a new technician's comments.
- More experienced techs can double-check their interpretation
- Experienced technicians can “auto-comment” their gels before doing manual assessment
- (Future:) Clients will be able to get a more complete interpretation of their chromatograms.

There are many important ways that AutoCommenting can be used profitably to streamline a Core's data processing.

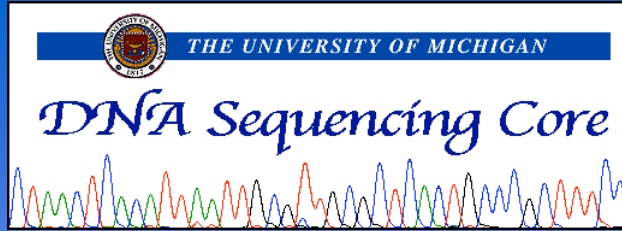
Automatic Chromatogram Commenting:

- *Very powerful diagnostic information can be derived from AutoComments*

FillStats & Co: a set of utilities that ...

- Automatically generates QA/diagnostics on all lanes
- Provides me with a quick look-up for quality as a function of sequencers, technicians, clients, etc.
- Warns us of specific failure modes (bleed-through and carry-over)
- Shows views from single samples through tens, hundreds, thousands or millions of samples.

Summary: most of the features I have described revolve around (I) the FillStats program, which generates Q/A statistics on all lanes of data we produce; and (ii) numerous web-based (CGI) scripts that display that statistical data to enhance our ability to comprehend the huge amount of data we produce.



Thanks to the Staff of the UM DNA Sequencing Core:

**Connie Esposito
Ellen Pedersen
Shamar Herron
Norman Roth**

**Suzanne Genik
Kayce Caugh
Christine Brannan
Jeff Longton**

Thank you to the ABRF for inviting me to speak, and that you to my staff, without whom I would be lost.



Final comment: spam is killing email, and companies spam because it is profitable to do so. In the future, you will see more and more mainstream companies intruding into your email box. DON'T sit back and allow it: complain loudly, and DON'T DO BUSINESS WITH COMPANIES THAT SPAM. For more information, please see:

<http://seqcore.brcf.med.umich.edu/cgi-bin/antispam.pl>