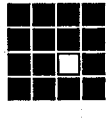
Data Coordinating Center
2Q3 Report



Wellcome Trust - AstraZeneca PLC - Bayer AG - Bristol-Myers Squibb Co - F.Hoffmann-LaRoche

THE SNP CONSORTIUM LTD.

Glaxo Wellcome PLC - Hoechst Marion Roussell AG - Novartis - Pfizer Inc. - Searle - SmithKline Beecham PLC

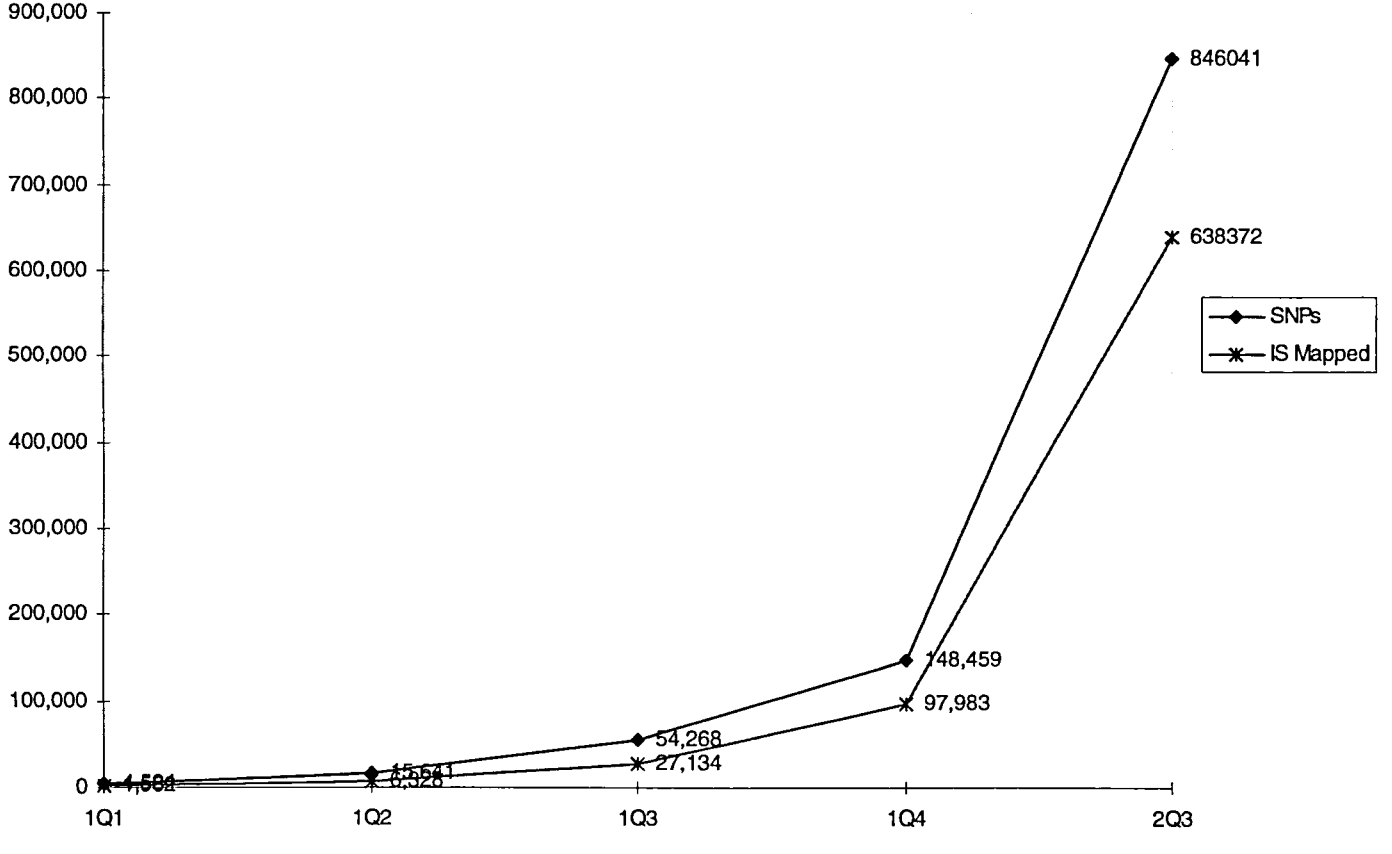
SNP Disposition 10/28/2000

Center	SNPs	Traces	Trace/SNP
Sanger	184,795	1,157,919	**6**
Wash U	163,283	1,686,393	**10**
Whitehead	497,963	2,567,241	**5**
		433,040	*new sanger*
TOTAL	**846,041**	**5,844,593**	
	81,493	*unprocessed Sanger*	
	200,000	*anticipated Wash U*	
	1,127,534	expected total	

SNP Mapping

Mapping to Genbank:			*Remaining*
	start	846,041	846,041
	duplicates	44,963	801,078
	no matches	23,565	777,513
	apparent repeats	139,141	638,372
	released	**638,372**	
Mapping to Golden Path			
	no match on GP	39,325	599,047
	version skew	23,621	575,426
	mapped	**575,426**	
Merge with NCBI			
	non-TSC SNPs	767,321	
	duplicates (38%)	217,396	
	unique SNPS in merge	**1,125,351**	

Cumulative Progress



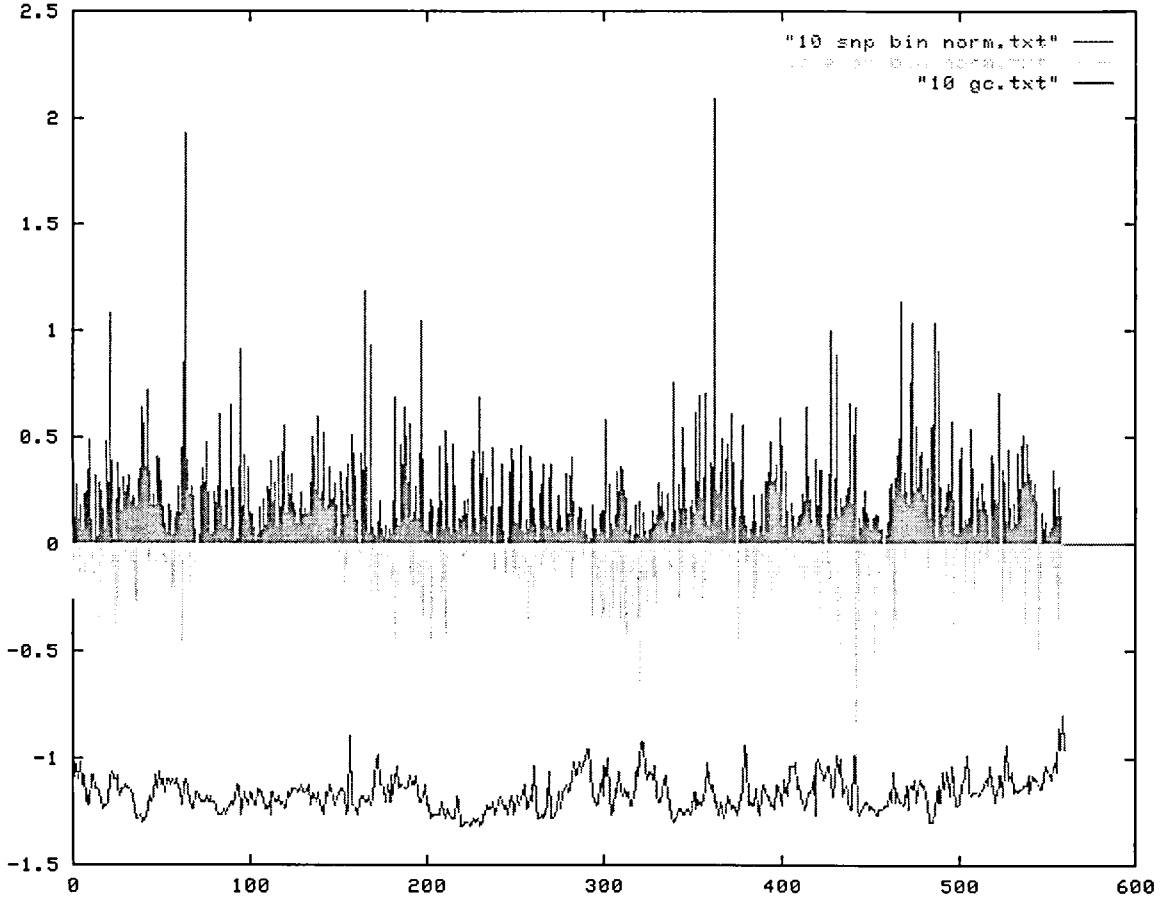
Chromosomal Distribution (R5)

Chrom	SNPs	TSC	GP Length	kb/SNP whole	kb/SNP TSC
1	107,431	48,290	242,664,353	2.26	5.03
2	83,124	53,614	250,687,854	3.02	4.68
3	76,365	43,837	218,037,158	2.86	4.97
4	59,737	40,406	184,288,662	3.09	4.56
5	102,182	43,340	199,630,150	1.95	4.61
6	82,813	36,507	189,279,554	2.29	5.18
7	59,132	28,169	162,512,756	2.75	5.77
8	46,298	30,152	148,977,423	3.22	4.94
9	49,270	27,931	117,862,627	2.39	4.22
10	44,855	23,626	140,316,765	3.13	5.94
11	68,811	28,208	148,783,472	2.16	5.27
12	48,711	25,497	143,786,455	2.95	5.64
13	43,650	24,497	101,873,169	2.33	4.16
14	30,412	13,804	98,454,640	3.24	7.13
15	29,188	16,503	85,695,754	2.94	5.19
16	31,261	15,522	84,459,027	2.70	5.44
17	27,229	11,635	85,438,480	3.14	7.34
18	36,587	16,810	85,988,934	2.35	5.12
19	21,715	6,289	69,273,146	3.19	11.01
20	25,787	12,145	72,153,882	2.80	5.94
21	7,370	7,030	65,454,843	8.88	9.31
22	12,146	7,698	45,617,351	3.76	5.93
X	27,146	12,607	176,974,140	6.52	14.04
Y	4,129	1,307	24,713,079	5.99	18.91
NA	957	957	32,747,882		
UL	6,128	1607	54,925,998		

LEN = 3,230,597,554 bp

Source	SNPs	Density
TSC	575,426	1 SNP/5.5 kb
dbSNP	549,925	1 SNP/5.3 kb
TOTAL	1,125,351	1 SNP/2.7 kb

SNP Distribution



Distribution of Gaps

Interval	% Intervals	Interval	% Intervals
0-1000	49.9	10k-11k	89.1
1000-2000	60.2	11k-12k	90.3
2000-3000	67.1	12k-13k	91.2
3000-4000	72.5	13k-14k	92.1
4000-5000	76.5	14k-15k	92.9
5000-6000	79.7	15k-16k	93.6
6000-7000	82.3	16k-17k	94.2
7000-8000	84.4	17k-18k	94.7
8000-9000	86.2	18k-19k	95.3
9000-10k	87.7	19k-20k	95.7

SNP Distribution, by Interval

%age intervals occupied by 1 or more SNPs (finished)

Interval	TSC (%)	dbSNP (%)
1000	12.5	25.2
2000	21.5	37.4
5000	40.3	56.4
10000	58.3	71.5
15000	68.7	79.7
20000	74.9	84.2
40000	85.7	91.4

Allele Distribution

Type	Count	Proportion
GA	197090	0.329
CT	198765	0.332
CA	52390	0.088
GT	53528	0.089
GC	51429	0.086
AT	44946	0.075
CGT	61	0.000
ACT	55	0.000
ACG	50	0.000
AGT	33	0.000
ACGT	23	0.000

Occupied Genes

RefSeq genes with at least 1 nearby TSC SNP*

SNPs	Genes	%	SNPs	Genes	%
0	635	22.5	24	12	0.4
1	377	13.3	25	9	0.3
2	315	11.1	26	9	0.3
3	283	10.0	27	5	0.2
4	196	6.9	28	4	0.1
5	175	6.2	29	2	0.1
6	146	5.2	30	1	0.0
7	93	3.3	31	3	0.1
8	78	2.8	32	4	0.1
9	68	2.4	33	2	0.1
10	54	1.9	34	2	0.1
11	60	2.1	37	2	0.1
12	34	1.2	38	1	0.0
13	34	1.2	39	1	0.0
14	24	0.8	41	2	0.1
15	29	1.0	42	1	0.0
16	17	0.6	43	1	0.0
17	16	0.6	44	1	0.0
18	9	0.3	45	2	0.1
19	6	0.2	46	6	0.2
20	18	0.6	47	1	0.0
21	6	0.2	48	6	0.2
22	3	0.1	49	4	0.1
23	4	0.1			

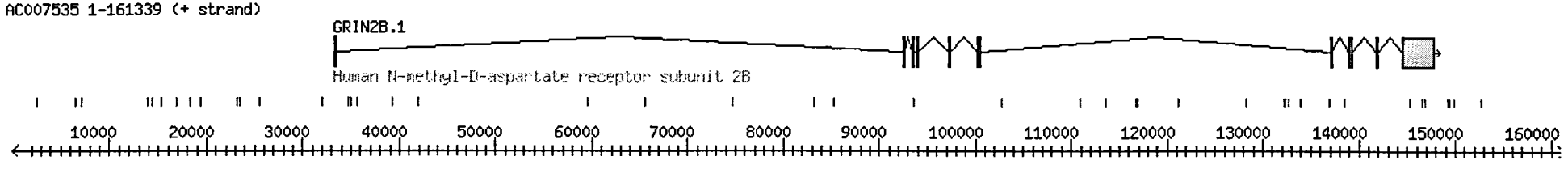
*between 5 kb upstream of 1st exon and 5 kb downstream of last exon

Distribution Among Genes

	Exons		Genes		Density
	in	near	in	near	
TSC	3%	57%	57%	78%	1 SNP/6.7 kb
dbSNP	12%	81%	78%	94%	1 SNP/1.1 kb

Chance that a RefSeq gene will have a nearby SNP

A “Typical” Gene

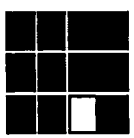


snp.cshl.org

- Release 5: 638,372 SNPs
- Mirror of dbSNP
- Everything in Golden Path coordinates
- RefSeq gene alignments
- Ensembl/Neomorphic gene predictions
- New interface to go live mid-November

The SNP Consortium

SMC Comments

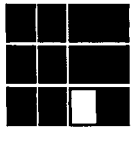


The
SNP Consortium Ltd.

10/31/00

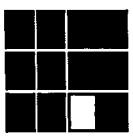
TSC Meetings--Confidential Disclosure

- Outlined in TSC Agreements.
- Protocol:
  - *Up front disclosure notification [receipt is elective].*
  - *Disclosure must be in writing or oral, followed by written documentation.*
- *Normal exclusions [e.g. prior possession, in public domain]*
- *Term: 5 years post research program completion*



SNP Map Project-Key Objectives

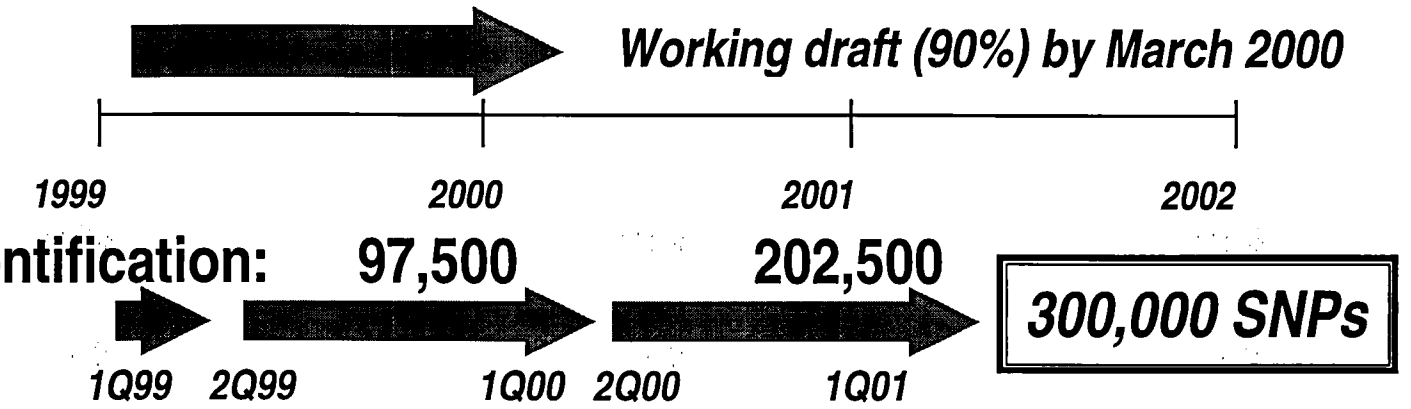
- Create the highest quality SNP map available
- Identified **300,000 SNPs** [1,000,000 SNP target]
- Map at least **150,000 SNPs** [800,000 SNP target]
- Broad, evenly spaced well annotated map.
- **Two year** production period--**5.3M reads** [8.4M target]
- Maximize public accessibility >> effective IP mgnt [550,000 "unencumbered" SNP target]
- $**45M** -Budget



Scientific Plan

Objectives and Timetable:

Human Genome Project:



The SNP Consortium Ltd.

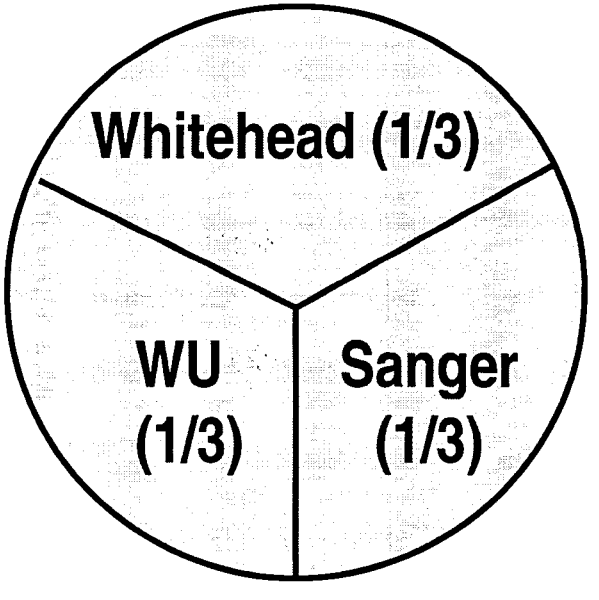
10/31/00

Metrics: Measuring the Productivity & Quality (I)

SNP Identification

- **Metrics:**

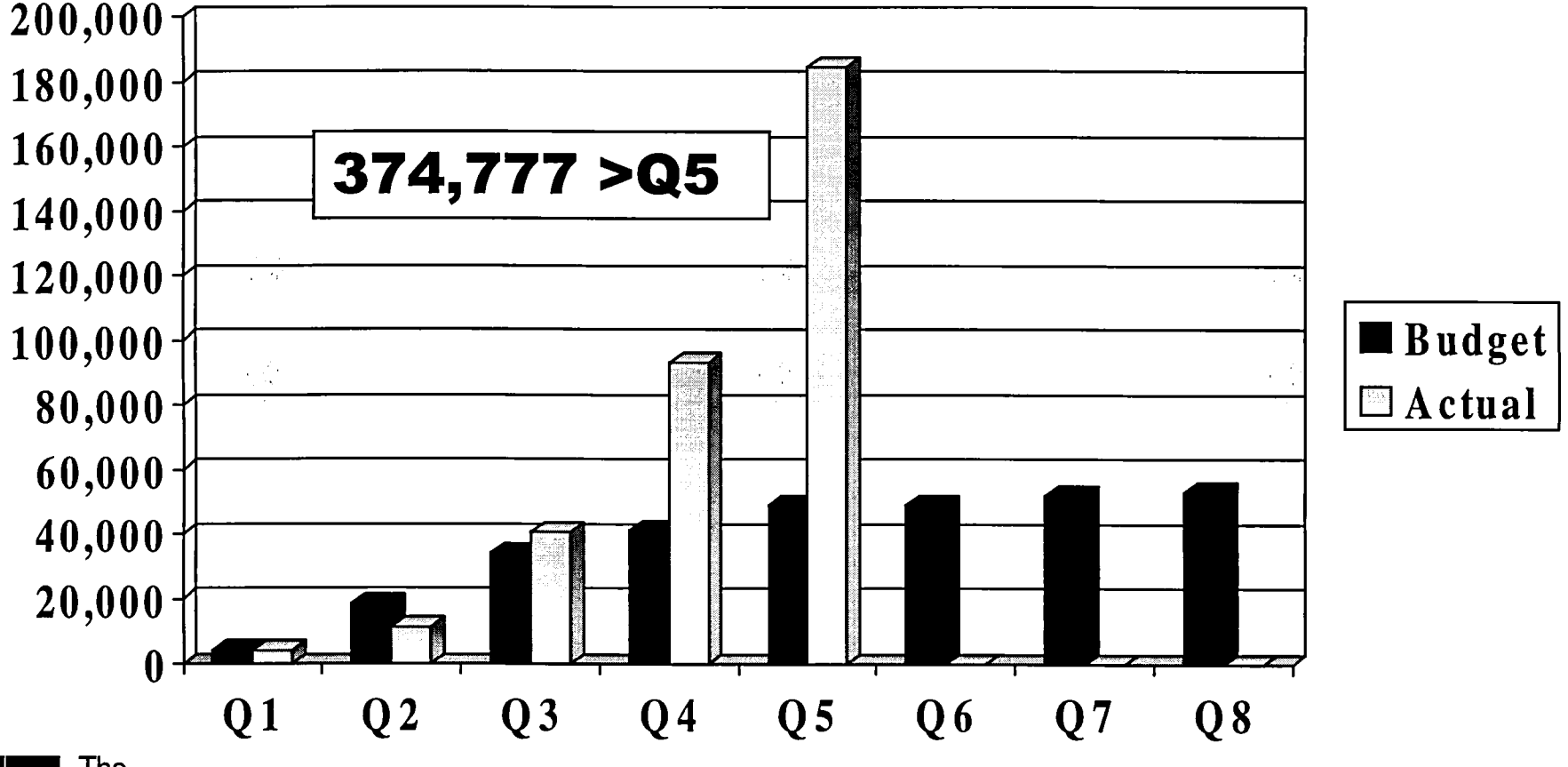
Quarter	SNPs Identified	
Q1 99	*Pilot*	
Year 1:		
Q2 99	*4,000*	
Q3 99	*18,500*	
Q4 99	*33,750*	
Q1 00	*41,250*	97,500
Year 2:		
Q2 00	*48,750*	
Q3 00	*48,750*	
Q4 00	*51,750*	
Q1 01	*53,250*	202,500
Total		**300,000**



- **Quality: ~ 95% accuracy**

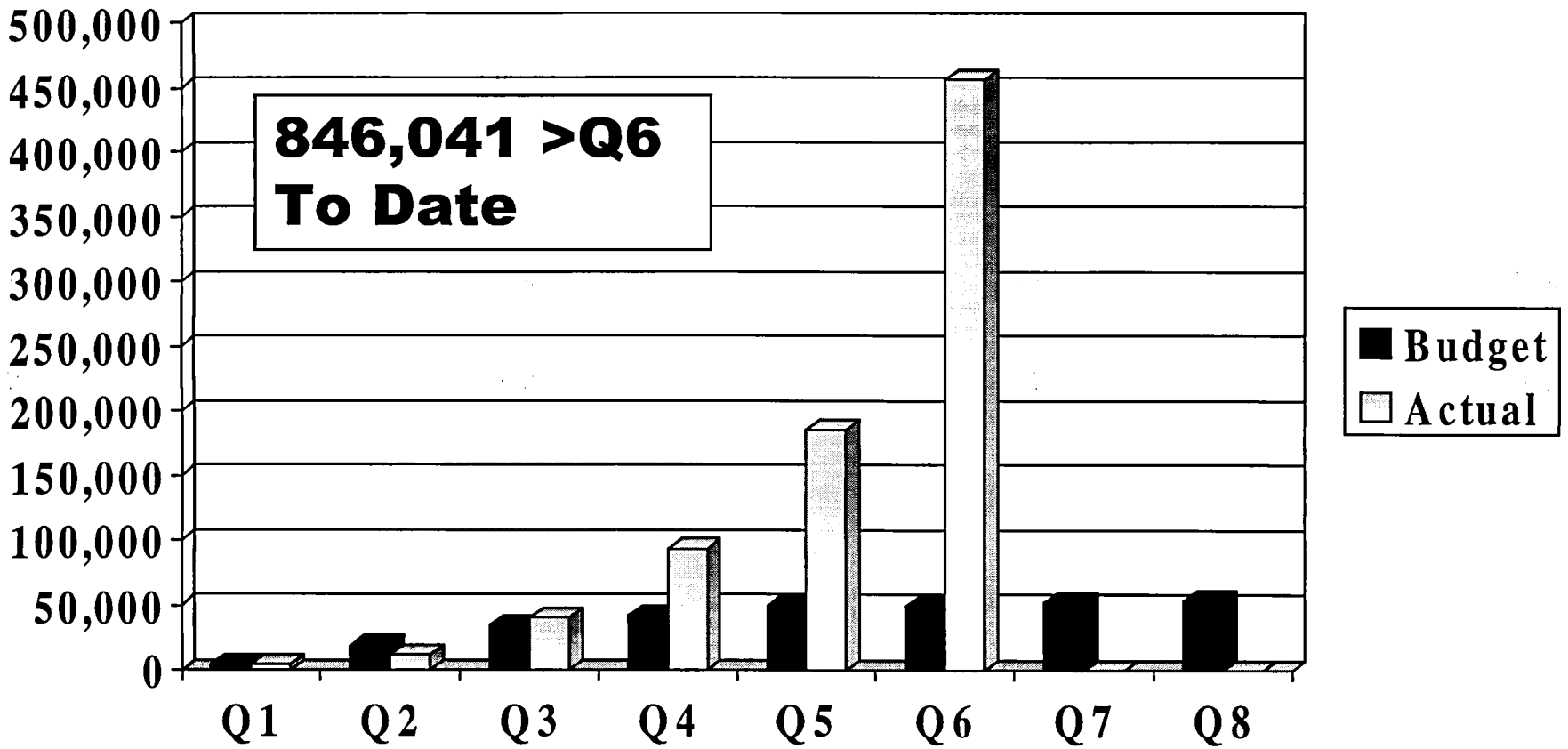
Program Productivity Profile *SNP ID*

Quarterly SNP ID Production--Budget vs Actual



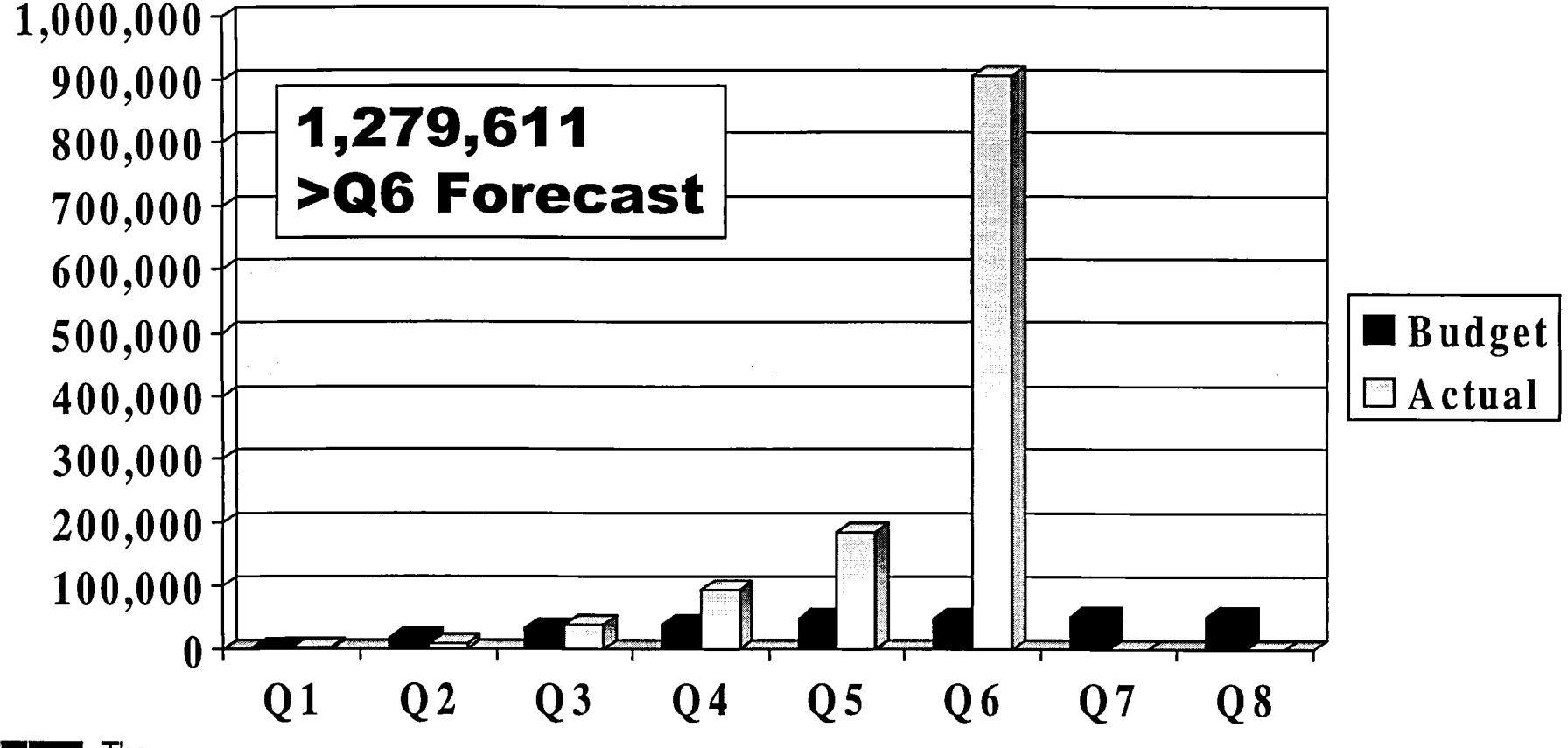
Program Productivity Profile *SNP ID to Date*

Quarterly SNP ID Production--Budget vs Actual



Program Productivity Profile *SNP ID 11/00*

Quarterly SNP ID Production--Budget vs Actual

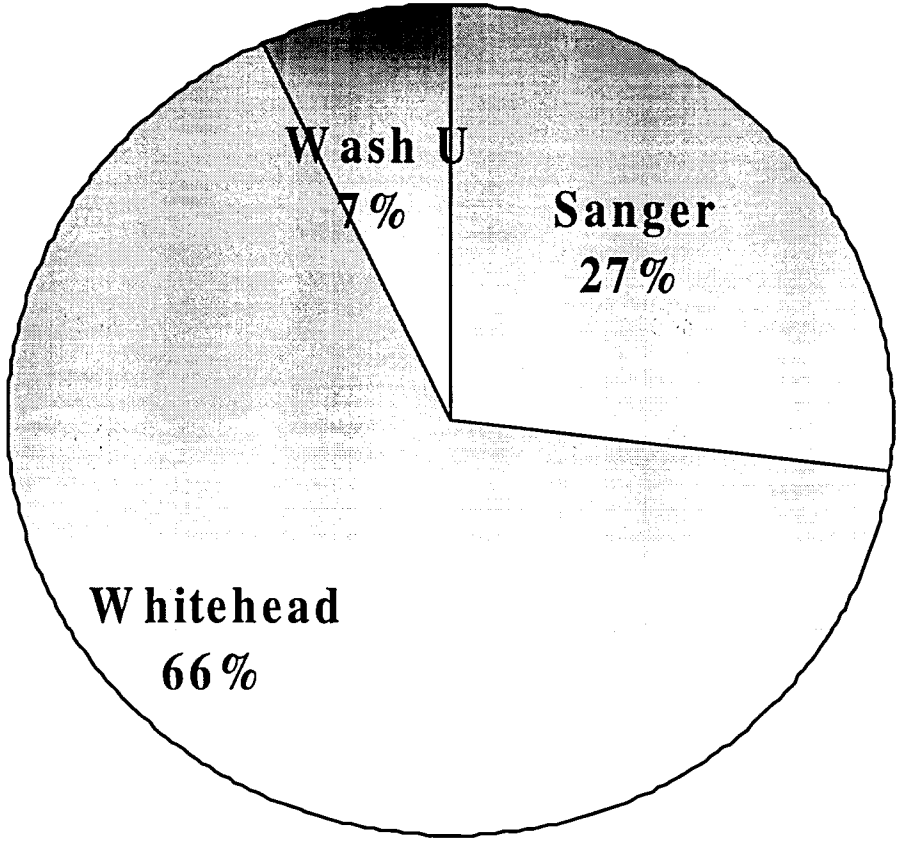


The SNP Consortium Ltd.

10/31/00

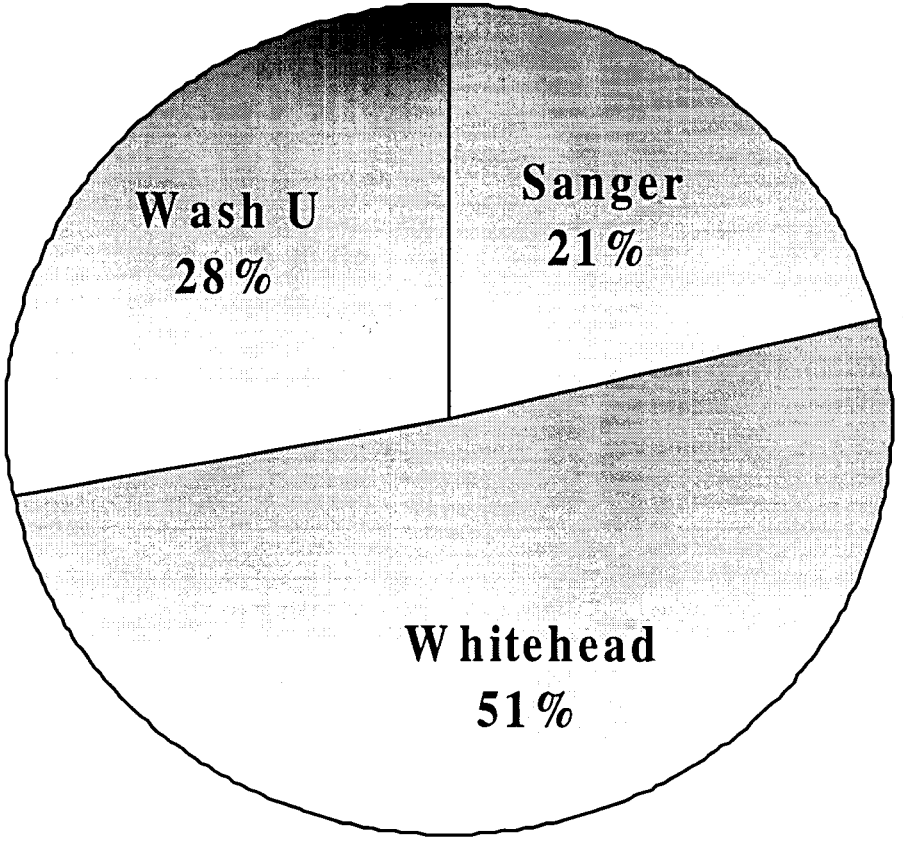
Program Productivity Profile

SNP Identification Production Mix—Processed to date by the DCC



Program Productivity Profile

SNP Identification Production Mix—Program Total F

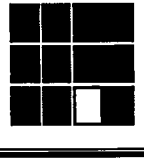


Metrics: Measuring Our Production to Date

SNP Identification – Total Program F [with NHGRI $s]

Quarter	SNPs Plan	SNPs Actual*	Variance to Plan
Q1 99	*Pilot*		
Year 1:			
Q2 99	*4,000*	4,519	+519
Q3 99	*18,500*	11,230	-7,270
Q4 99	*33,750*	40,368	+6,618
Q1 00	*41,250*	92,342	+51,092
	97,500	148,459	+50,959
Year 2:			
Q2 00	*48,750*	226,323	+128,823
Q3 00	*48,750*	904,827	+807,327
Q4 00	*51,750*		
Q1 01	*53,250*		
	202,500	1,002,327	+897,327
Total PTD	300,000	*1,279,611*	*+979,611*

+327% to plan



The SNP Consortium Ltd.

10/31/00

10/30/00	TSC SNP Production						Actual/Forecast	
	Q1	Q2	Q3	Q4	Q5-6	Q7-8		
	Q2`99	Q3`99	Q4`99	Q1`00	Q2`00	Q2`00		
Aggregate SNP ID Target/Quarter	4,000	18,500	33,750	41,250	97,500	105,000	300,000	100,000
Cummulative Output		22,500	56,250	97,500	195,000	300,000		
Minimum--SNP ID Requirement @ 90%							270,000	
Budgeted SNP Production								
Sanger Center	1,333	6,167	11,250	13,750	32,500	*35,000*	100,000	
Wash University	1,333	6,167	11,250	13,750	32,500	*35,000*	100,000	
Whitehead	1,333	6,167	11,250	13,750	32,500	*35,000*	100,000	
Quarterly Budget--SNPs ID	4,000	18,500	33,750	*41,250*	*97,500*	*105,000*	300,000	
Actual/Forecast SNP Production								
Sanger Center [Included Genomic SNPs]	1,340	-	17,494	48,855	116,396	81,493	265,578	- 21%
Wash University	1,456	4,759	5,628	21,755	34,435	*296,000*	*364,033*	28%
Whitehead	1,723	6,471	17,246	21,727	75,499	527,334	650,000	51%
Quarterly Actual--SNPs ID	4,519	11,230	40,368	92,337	226,330	904,827	1,279,611	
CUM YTD--SNPs ID	4,519	15,749	56,117	148,454	374,784	1,279,611		
CUM Program--SNPs ID Budget Variance	519	(6,751)	(133)	51,087	128,830	799,827	979,611	
Variance to Budget								
Sanger Center [Included Genomic SNPs]	7	(6,167)	6,244	35,105	83,896	46,493	165,578	17%
Washington University	123	(1,408)	(5,622)	8,005	1,935	*261,000*	*264,033*	27%
Whitehead	390	304	5,996	7,977	42,999	492,334	550,000	56%
Quarterly Variance to Budget--SNPs ID	519	(7,270)	6,618	51,087	128,830	799,827	979,611	
SNP Mapping--B vs F with the hedge program								
Direct RH mapping--Budget		1,000	8,000	9,000			18,000	
In Silico--Cumulative--Budget				29,250	78,000	132,000	132,000	
Total Mapped SNPs--Budget				30%	40%	44%	150,000	
Actual/Forecast SNP Mapping								
Net Direct RH mapping--Stanford		-	-	1,346	2,029	2,029	2,029	
In Silico--Cumulative	1,499	4,725	26,998	101,370	297,645	966,106	966,106	
In Silico--Mapping Rate/Working Draft	33%	30%	48%	68%	*79.4%*	*76%*		
Total Mapped SNPs	1,499	4,725	26,998	102,716	299,674	968,135	968,135	
Variance to Budget--SNPs Mapping				*69.2%*	*80.0%*	*75.7%*	818,135	

SMC--Key "Outstanding Follow-Up Items"

- Quality Management Results—Orchid Summary
- TSC Member Inputs to LS on TSC Website
- "Finishing Plan">>Q7

Orchid – TSC AF program

Original Proposal

- 60,000 SNP assays developed
- Genotypes from 44 samples from 3 ethnic groups
- Report allele frequency

Current Proposal

- Assay design 100,000 SNPs
- Select and bin SNPs into 3 types
  - Disease "Hot Spots"
  - SNPs in and near all known ADME genes
  - Remainder randomly dispersed across genome
- Genotype on following samples
  - CEPH families back to grandparents to analyze chromosome phase and generate haplotypes and allele freq.
  - African American and Asian populations for allele freq.

Orchid – TSC AF program

Original Budget

Reagents	**3.6MM**
primers	
PCR	
SNP-IT	
Coriell cell-lines	
Labor (fully loaded)	**2.6MM**
assay design	
geontyping	
informatics	
Total	**6.2MM**

Current Budget

Reagents	**5.3MM**
primers	
PCR	
SNP-IT	
Coriell cell-lines	
Labor (fully loaded)	**3.7MM**
assay design	
genotyping	
informatics	
Total	**9.0MM**

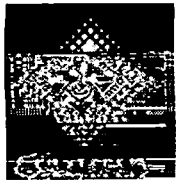
Difference $2.8MM

Reasons for Variance

- More SNPs assessed in assay design stage
- Assay design for Kwok lab
- Additional informatics to bin SNPs
- Bulk reagent pricing not as beneficial as forecasted
- More samples being analyzed

Benefits of Adjustments

- Allele frequency on better set of SNPs
- Chromosome phasing (haplotyping) on 60,000 SNPs
- Primer designs and reagent sets validated for genotyping on variety of platforms
  - SNPstream 25K
  - Affymetrix (GeneFlex) chips
  - SNPstream 5K (Luminex flow cytometry)
  - Sequencers
  - Others to follow…
- PCR primer designs on 60,000 SNPs validated



Orchid – TSC AF program

Original Proposal

- 60,000 SNP assays developed
- Genotypes from 44 samples from 3 ethnic groups
- Report allele frequency

Current Proposal

- Assay design 100,000 SNPs
- Select and bin SNPs into 3 types
  - Disease "Hot Spots"
  - SNPs in and near all known ADME genes
  - Remainder randomly dispersed across genome
- Genotype on following samples
  - CEPH families back to grandparents to analyze chromosome phase and generate haplotypes and allele freq.
  - African American and Asian populations for allele freq.

Orchid – TSC AF program

Original Budget

Reagents	**3.6MM**
primers	
PCR	
SNP-IT	
Coriell cell-lines	
Labor (fully loaded)	**2.6MM**
assay design	
geontyping	
informatics	
Total	**6.2MM**

Current Budget

Reagents	**5.3MM**
primers	
PCR	
SNP-IT	
Coriell cell-lines	
Labor (fully loaded)	**3.7MM**
assay design	
genotyping	
informatics	
Total	**9.0MM**

Difference $2.8MM

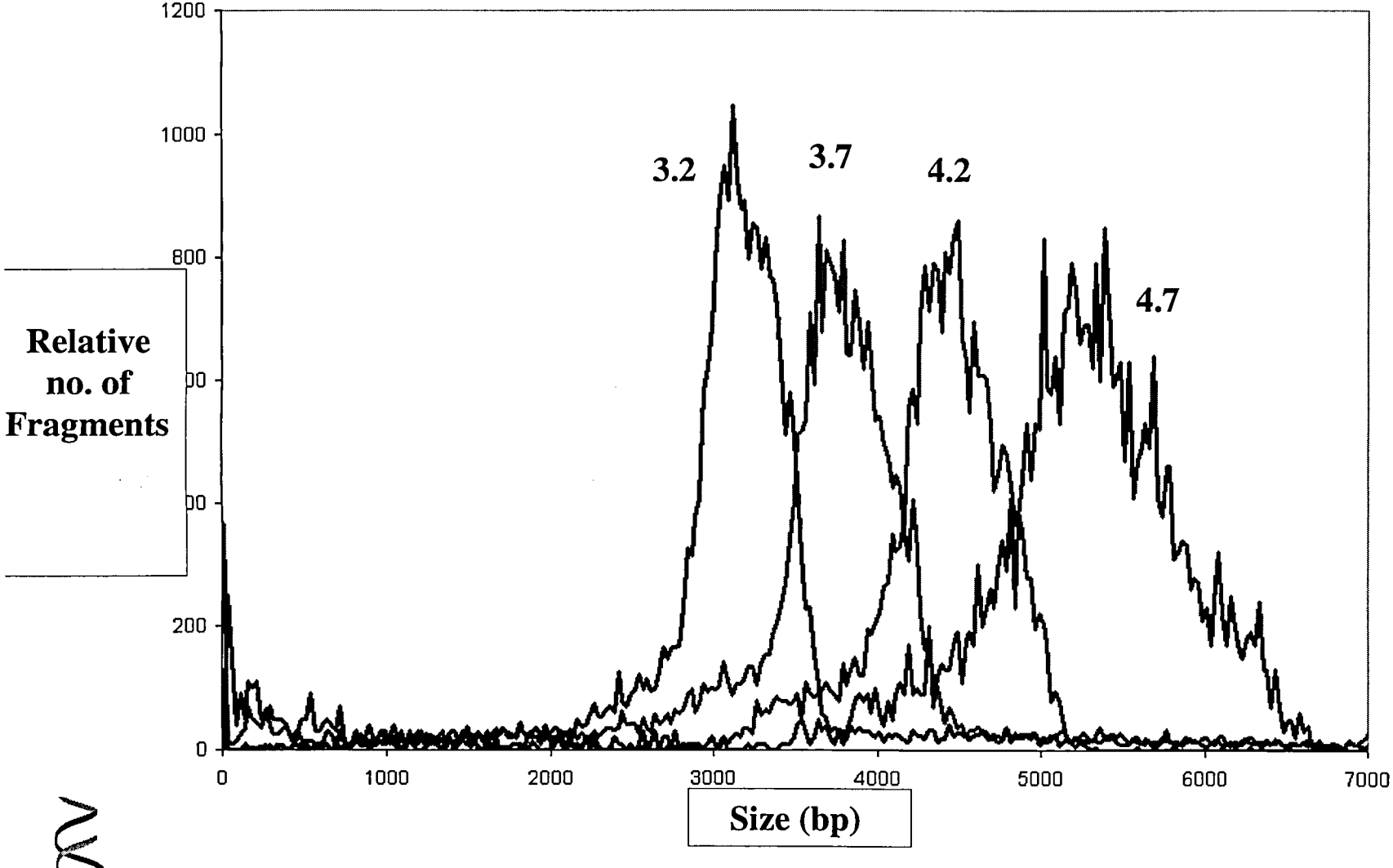
Summary – Q6 October 2000

- Sequencing & analysis of libraries
- Candidate SNP identification
- Verification
- Summary of Sanger Centre Programme

Production sequencing – Q6

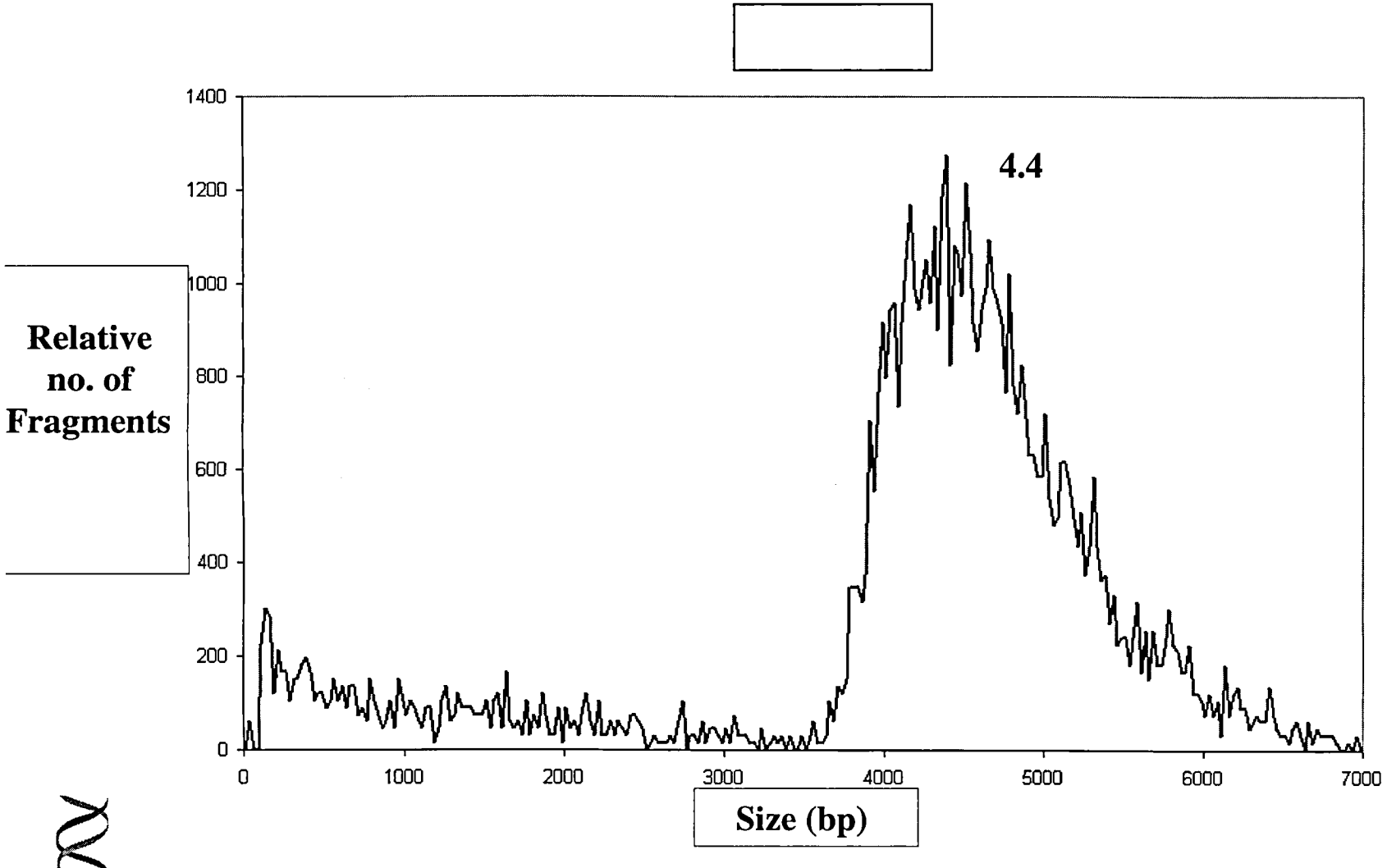
	Library				
	PvuII 3_2	PvuII4_2	PvuII4_7	RS1-2	Total
Attempted reads	65,344	203,183	158,190	179,608	606,325
Clipped reads	35,424	119,546	86,202	90,294	331,466
Repeat contamination	21,809	58,926	25,582	29,674	135,991
# good inserts	29,348	96,439	67,830	68,908	262,525
Yield of good inserts	45%	47%	43%	38%	43%

Fragment distribution - PvuII libraries



The Sanger Centre

Fragment distribution – RS1-2 library



The Sanger Centre

Candidate SNPs – Q6

Method	Library	Reads analysed	No. of SNPs	Reads/SNP
	PvuII			
GA (71%)	1.0+1.1+1.3	614k	123,387	5.0
GA (71%)	3.2+3.7+4.2+4.7	301k	81,241	3.7
	RS			
GA (69%)	4.4	69k	20,454	3.4

Total number of unique SNPs this quarter: 81,493

Summary of verification data

	Method	Verification (success/attempted)	
Pilot study:	**RRS**	**79/83**	**95.2%**
	GA-RRS (F)	**115/122**	**94.3%**
	GA-RanS (F)	**33/34**	**97.1%**
Main programme:	**RRS**	**122/126**	**96.8%**
	GA-RRS (UF)	**267/273**	**98%**
	GA Large Insert (UF)	**156/160**	**97.5%**
	Genomic Overlap (F/UF)	**(83% SNPs observed in panel of 24)**	

Q1-Q6 Summary

Study		Total SNPs	Unique SNPs
Pilot on chromosome 22:			
	RRS	743	
	GA-RRS (F)	1,362	
	GA-RanS (F)	1,920	3,444
Main programme:			
	RRS/PvuII	69,731	
	GA/PvuII (71%)	204,628	
	GA/RS1-2 (69%)	20,454	262,314
		Total:	**265,758**

The Sanger Centre

SsahaSNP

- SsahaSNP's performance was speeded up and can now process reads at 50/second (was 30/second).
- SsahaSNP has been used to process all of the TSC data, giving a common analysis platform. This has proven very useful in comparing SNPs from the different libraries used across the three sequencing centres.
- A detailed comparison was made between SsahaSNP and WUGSC/PolyBayes on the WUGSC data set, giving further insights into the strengths of both methods.