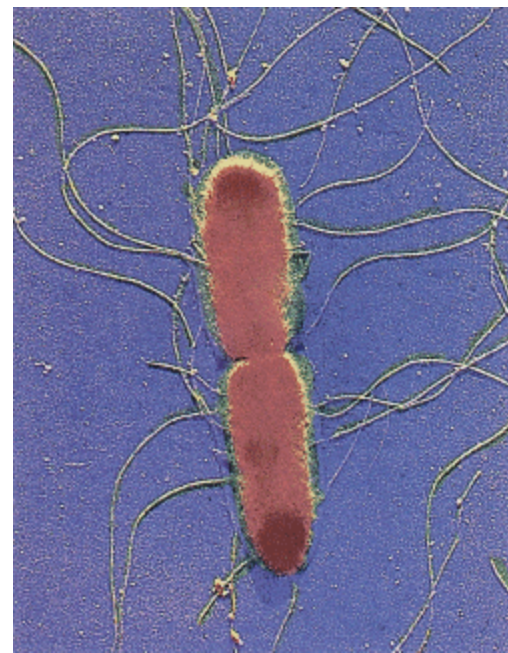


QuickTime™ and a  
None decompressor  
are needed to see this picture.

# State of the art sequencing - microbial applications

Julian Parkhill

Illumina user group meeting,  
Crete, April 2009



# WTSI - current installed machine base

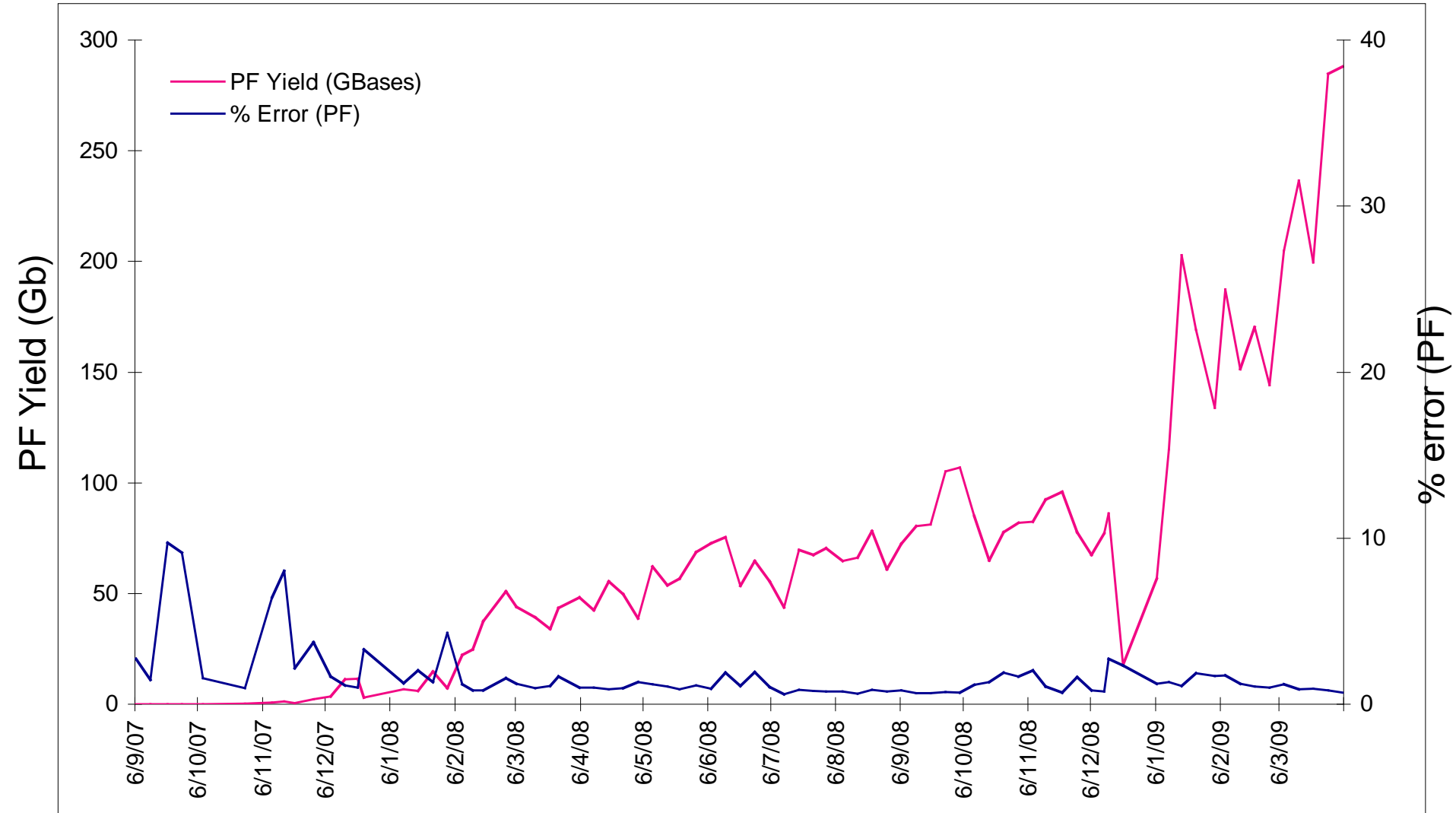
38 Illumina GAIIx

2 Roche 454FLX

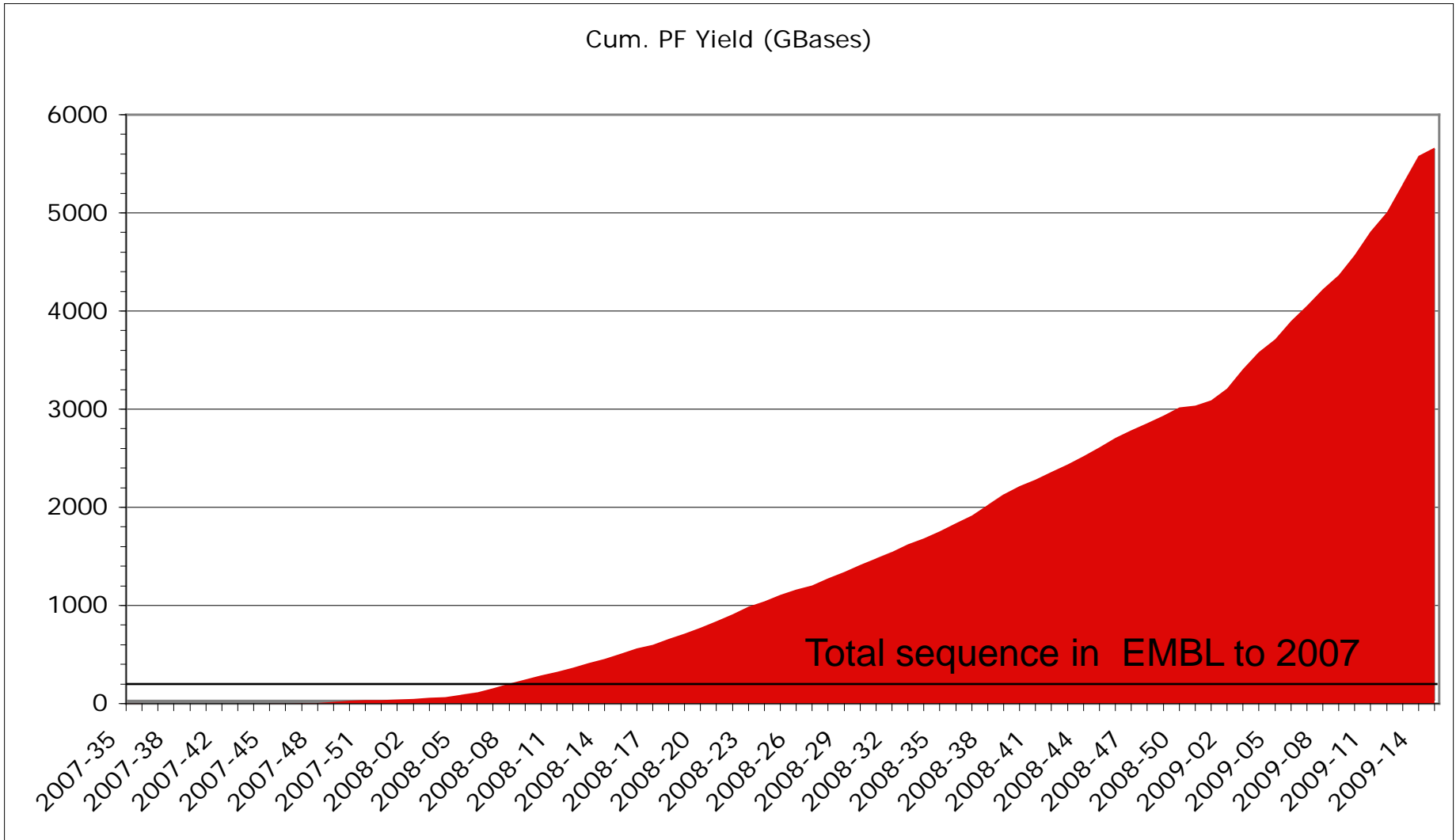
35 ABI 3730xl



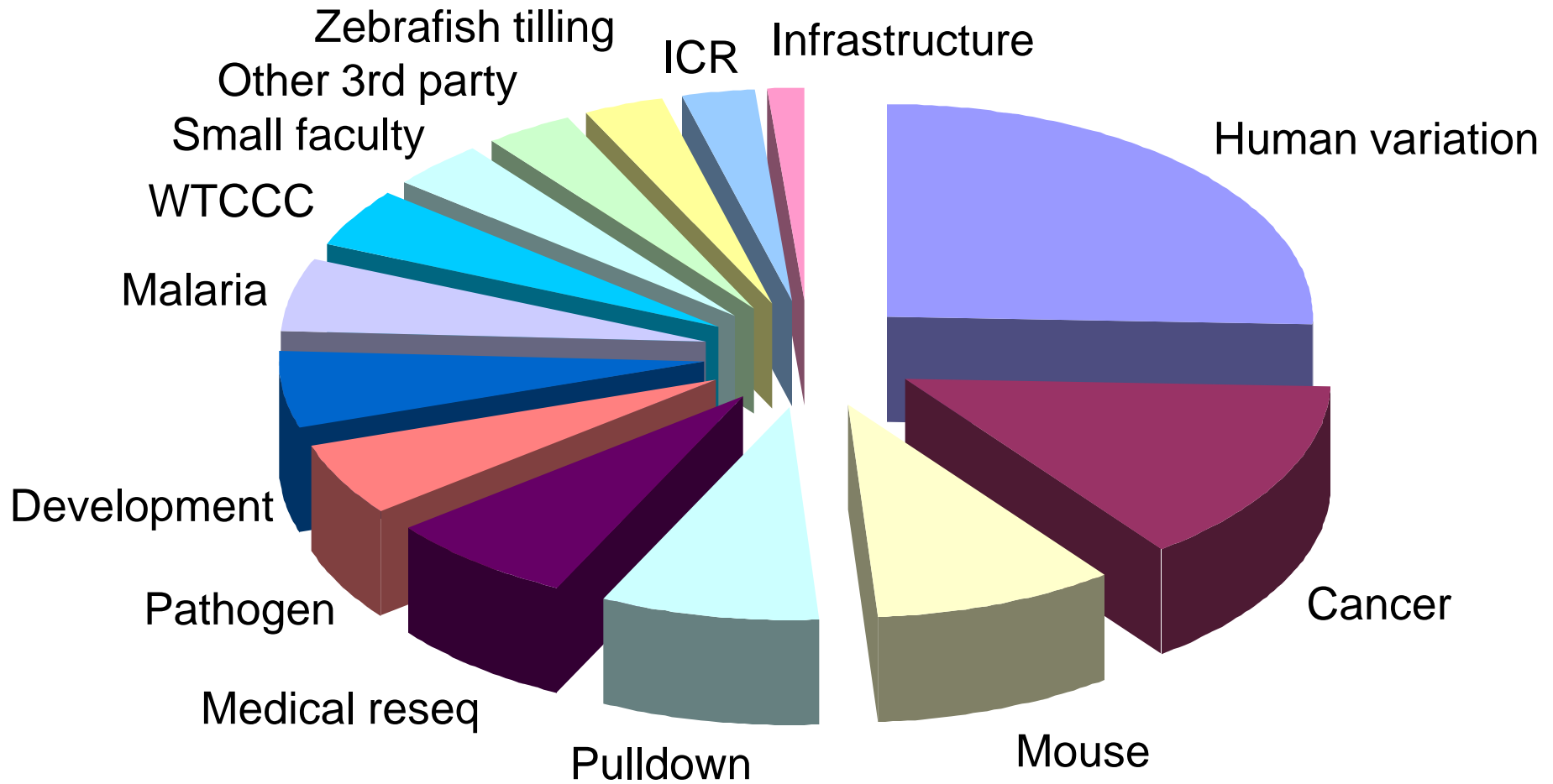
# WTSI - Illumina weekly throughput



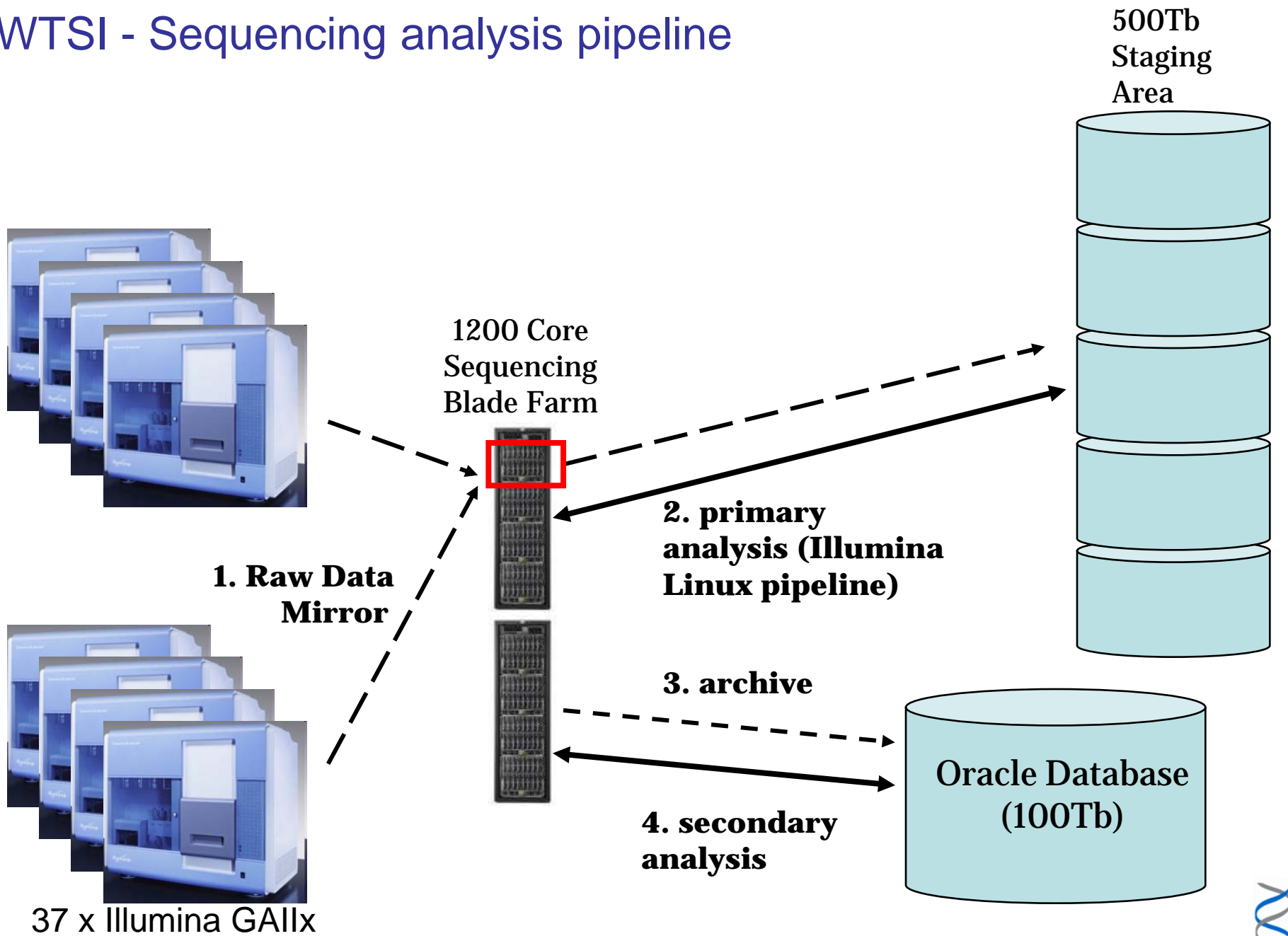
# WTSI - Illumina cumulative output



# WTSI - major project usage



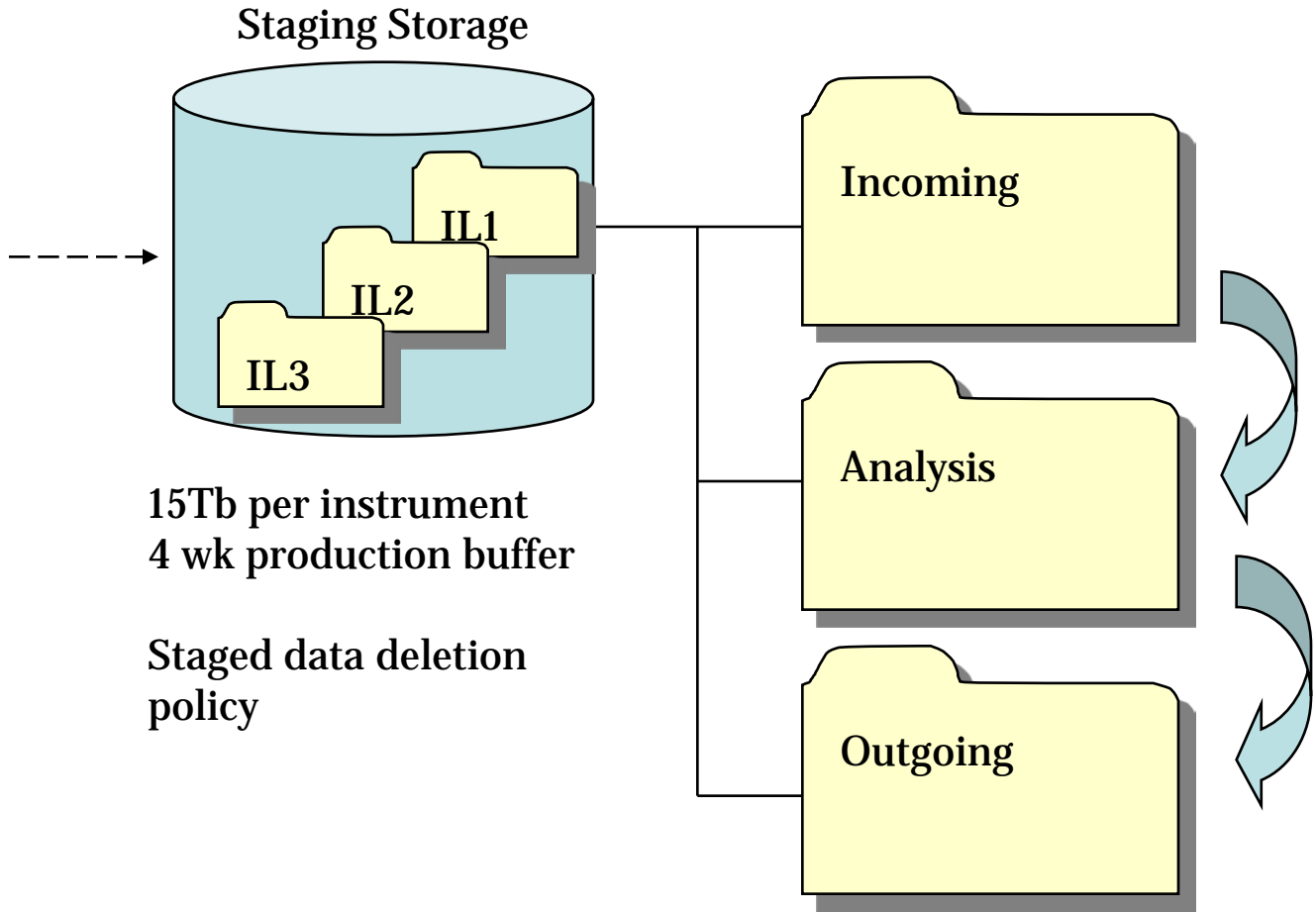
# WTSI - Sequencing analysis pipeline



# WTSI - Image data management



Wget/mirror



# WTSI - Data analysis and storage pipelines

- Data in data centre, not labs.
- 500Tb disk “staging” area to collect raw data.
- 4x100Tb Lustre file systems + 2x50Tb NFS partitions.
- No backup
- Each GA2 has notional 15Tb allocation of staging space.
- Enough for ~3-4 weeks production buffer. Not for much longer.
- Raw data mirrored (by simple wget/FTP) over dedicated (gigabit Ethernet) network as it is produced on to the staging area.
- Periodic on-instrument image deletion to maximize disk available in event of network outage. Only retain last 10 cycles of image data on instrument PC.
- Low resolution tile image store in Oracle for QC
- Data deletion policy gradually removes older data from staging area.
- Data management is almost completely automatic and resilient to network failures.
- Modular design means it is expandable



# WTSI - Data analysis and storage pipelines

- 1200 core IBM blade-based compute farm
  - primary image processing and secondary analysis/reference alignments. Mostly 8-core blades with 2Gb RAM/core
- All nodes see all staging storage file systems as local disk.
  - Lustre was the choice although recently this is not so clear. Very important to choice as I/O performance is crucial
  - Primary Illumina pipeline run under Linux and controlled by “make”.
- Pipeline analysis is parallelised across 8 dual-quad core blades (64 cores) using LSF (Platform Computing). Run processed <12 hours
- Sequence data (fastq format) and trace info (as SRF) stored as binary LOBs in Oracle database.
- FUSE layer exposes DB contents as a filesystem.
- Intend to use EBI ERA as long term off-site SRF archive from 2009.



# WTSI - recent modifications to molecular biology pipelines

Step	Modification	Benefit
Fragmentation	Accoustic shearing	Greater proportion of DNA in desired size range
Post-fragmentation	Double size selection	Fewer chimeric templates
Ligation	Ubiquitous use of PE oligos	Convenience and flexibility
PE size selection	Thin gel slice	Improved robustness
Gel extraction	Cold dissolution of gel slices	Decrease in GC bias

Note: these modifications are used as and when required, not all for every sample



# WTSI - recent modifications to molecular biology pipelines

Step	Modification	Benefit
PCR	optimised template quantity	cleaner libraries, fewer PCR duplicates
	optimised PCR conditions	improved yield, fewer cycles of amplification
	SPRI beads	fewer adapter dimers
	PCR-free sequencing	removal of duplicates; decrease in GC bias
	direct sequencing of short amplicons	reduced amplification bias



# WTSI - recent modifications to molecular biology pipelines

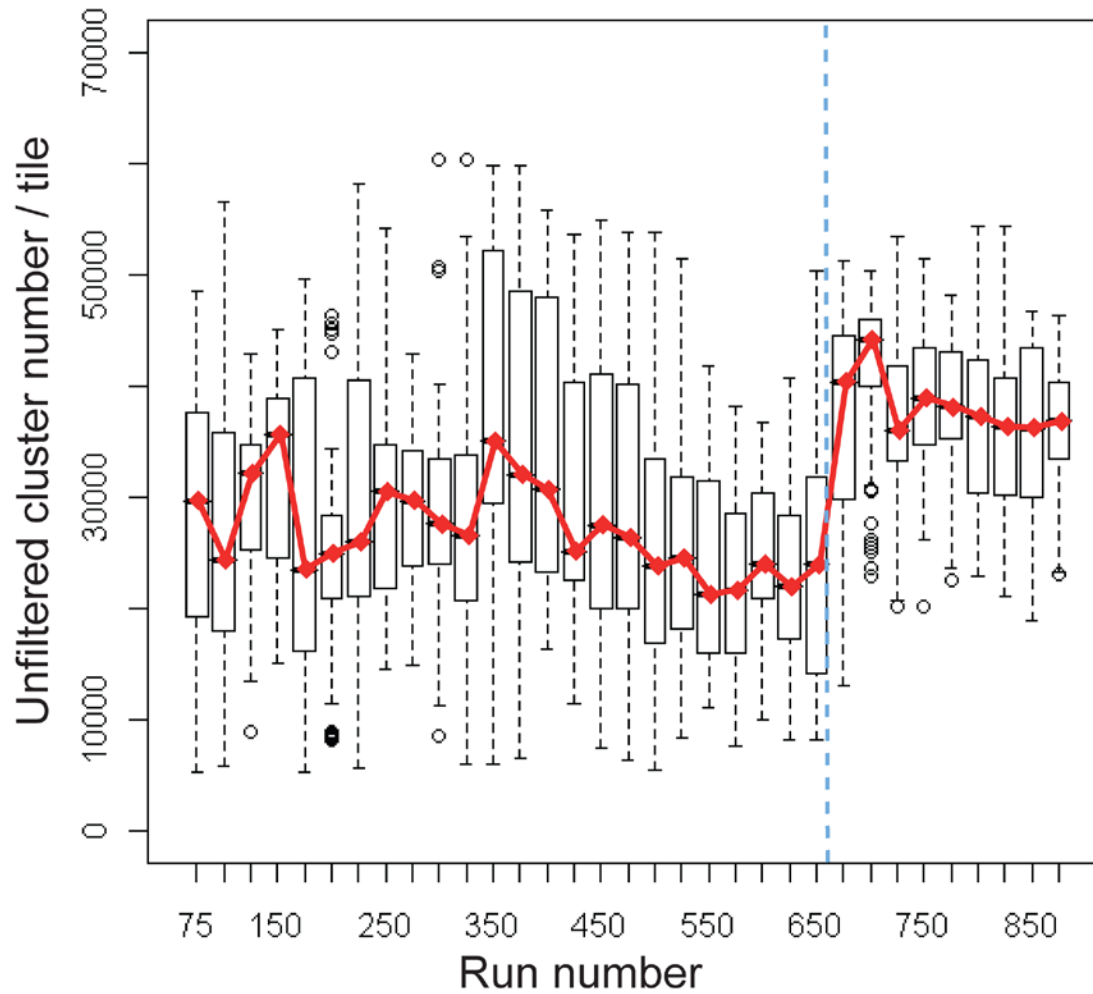
Step	Modification	Benefit
Quantification	qPCR assay	Accurate prediction of cluster density
Denaturation	Modified hybridisation buffers	Counteracts pipetting errors and allows sequencing of dilute templates
Amplification	QC step	Allows verification of cluster amplification and density

Note: We have modified buffers further since the paper as the recipes given gave poor results with the GA2 flowcells.



# WTSI - recent modifications to molecular biology pipelines

Use of qPCR for quantification



# A large genome center's improvements to the Illumina sequencing system

Michael A Quail, Iwanka Kozarewa, Frances Smith, Aylwyn Scally, Philip J Stephens, Richard Durbin, Harold Swerdlow & Daniel J Turner

**NATURE METHODS** | VOL.5 NO.12 | DECEMBER 2008 | 1005

## Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes

Iwanka Kozarewa<sup>1,2</sup>, Zemin Ning<sup>1,2</sup>, Michael A Quail<sup>1</sup>, Mandy J Sanders<sup>1</sup>, Matthew Berriman<sup>1</sup> & Daniel J Turner<sup>1</sup>

**NATURE METHODS** | VOL.6 NO.4 | APRIL 2009 | 291

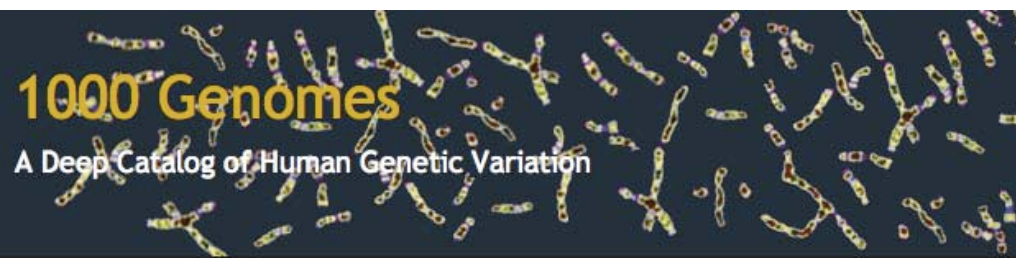
Please see poster from Iwanka Kozarewa on 96/lane indexing

Dan Turner - [djt@sanger.ac.uk](mailto:djt@sanger.ac.uk); Mike Quail - [mq1@sanger.ac.uk](mailto:mq1@sanger.ac.uk)



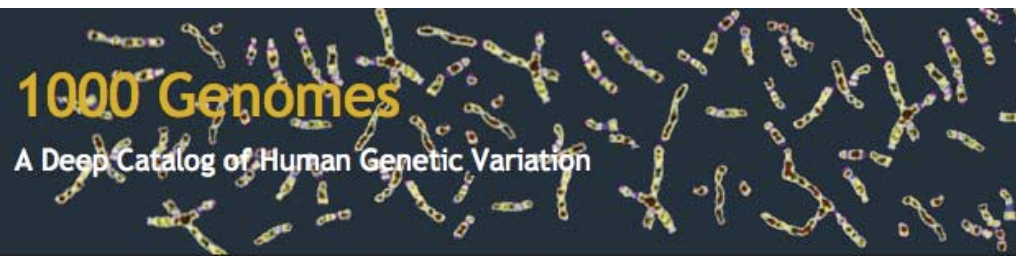
# WTSI - current projects - 1000 genomes

- International Consortium to generate a high resolution catalogue of human variation to support medical genetics
  - Production: Baylor, BGI, Broad, Sanger, WashU, Berlin + AB/Illumina/454
  - Analysis: + many statistical and population geneticists
  - Data Coordination: EBI, NCBI
  - Samples/ELSI: + expertise in ethics and population sampling
  - Funders: NHGRI, Wellcome Trust, Chinese, companies
- Produce an open resource building on HGP, HapMap etc.
  - Anonymised samples
  - Data publicly available, cell lines available



# WTSI - current projects - 1000 genomes primary goals

- Discover effectively all shared variation
  - e.g. any variant down to 1% minor allele frequency in a population in the accessible genome has a 95% chance of being identified
  - Structural variants as well as SNPs
  - Deeper in gene regions, down to 0.5% to 0.1% MAF
- Characterise allele frequency and haplotype context
  - Call and phase the variants on the sampled individuals
  - This will support tagging by genotyping and imputation based approaches
- Do this in multiple populations



# WTSI - current projects - 1000 genomes current status

Project started Feb 2008, >5Tb raw data already

- Three pilots during 2008
  - Pilot 1: 3x60 samples low coverage: Complete
    - European (CEU) 4x, African(YRI) 2x, East Asian (CHB/JPT) 2x Complete
  - Pilot 2: European and African trios at 20x
  - Pilot 3: 1000 genes in 1000 people ~50% data available
  - Develop and evaluate methods for data Simulations
  - collection and analysis Initial variant calls from pilots
- Main project Collection and consent structure established
  - Additional samples UK, Finland, Spain, S China cohorts approved
    - New samples beyond HapMap with appropriate ethics
  - Design to be finalized based on pilots

Project meeting at ASHG November 10/11: plan 3 x 400 people x 4x in 2009

Richard Durbin - rd@sanger.ac.uk



# WTSI - current projects - Cancer Genome Project

- Whole Cancer Genome Shotguns
  - Several cancers and matching normal DNAs to >30X coverage
  - Paired end, mixed insert size, long reads (100bp)
- Paired-end low coverage shotguns for rearrangements in cancer genomes
  - 37bp paired ends, long insert
  - multiple tumour types: breast, renal, CLL, MPD, myeloma, adenoic cystic carcinoma, chordoma, osteosarcoma, ovarian cancer.
  - Investigation of genetic heterogeneity in metastatic pancreatic adenocarcinoma
- Deep resequencing of therapeutic targets to understand genetics of treatment resistance
  - large amplicon resequencing
  - RNA-seq/whole transcriptome sequencing for:
    - some samples with low-coverage rearrangement shotguns
    - all samples with high-coverage whole genome shotgun



# WTSI - current projects - Mouse re-sequencing

## Sequencing of the genomes of 17 key mouse strains

**Backgrounds on which >5,000 knockouts have been made:**

*129S1, 129S5, 129P2*

**Common lab strains (progenitors of the collaborative/HS cross):**

*A/J, AKR/J, BALB/cJ, CBA/J, C3H/HeJ, DBA/2J, LP/J, NOD/LtJ, NZO/HILtJ, WSB/EiJ*

**Background strain for EUCOMM & KOMP gene targeting programmes:**

*C57/B6N*

**Wild-derived strains (Infection and cancer resistant):**

*PWK/PhJ, SPRETUS/EiJ and CAST/EiJ.*

**Initially 'guided' (mapped to reference) but ultimately *de novo* assemblies**



# Mouse re-sequencing strategy

## DNA from Jackson Labs and MRC Harwell

**Stage 1:** Release an ultra-high resolution map of SNPs and Indels by generating Illumina coverage of each strain.

- Release the data via dbSNP and Ensembl
- Genotype a subset of the SNPs

**Stage 2:** Generate a 'guided assembly' of each genome

- Release via Ensembl

**Stage 3:** Attempt complete *de novo* assembly

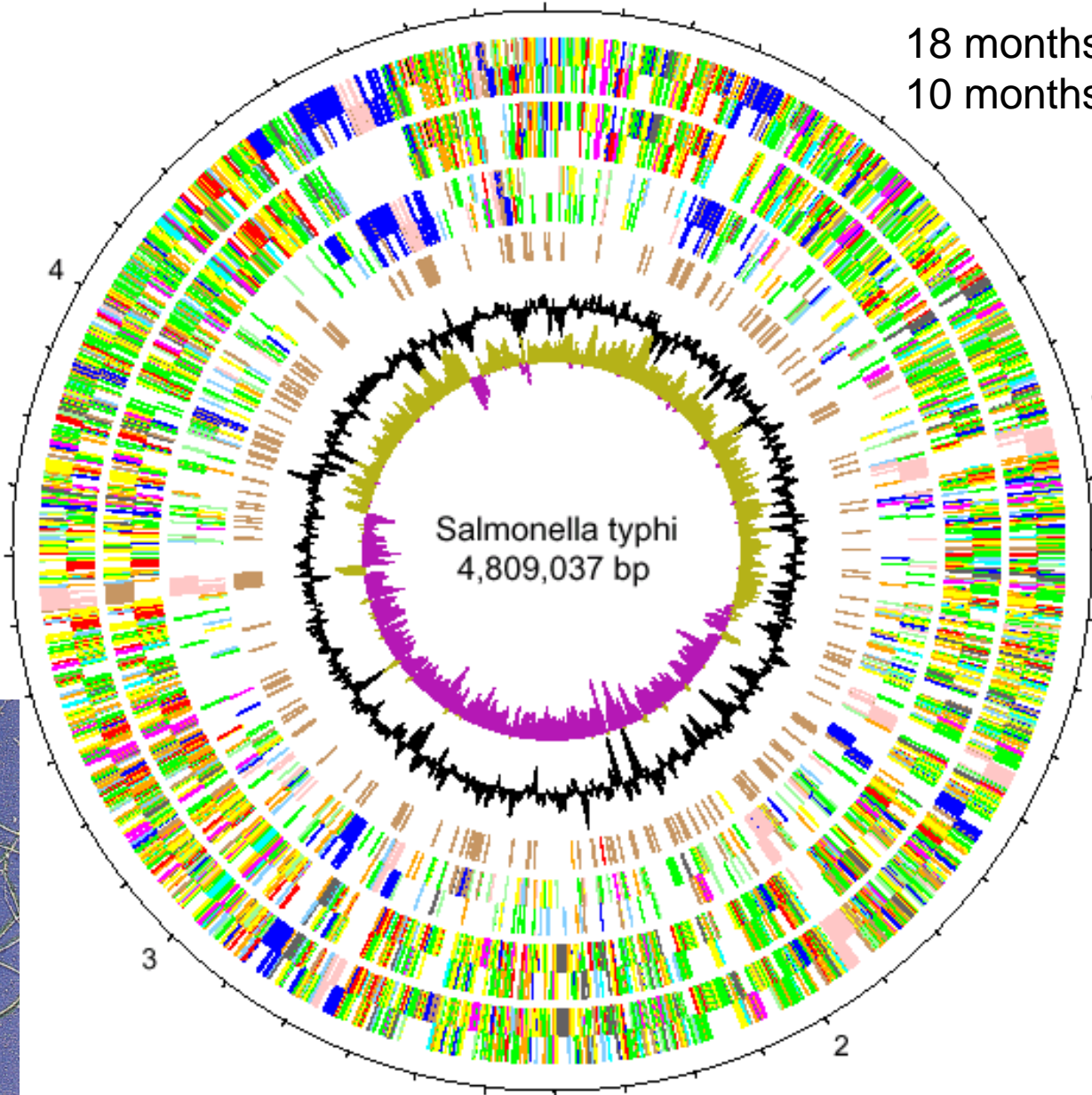
- Long read coverage may be required

**Progress: 20x coverage of NOD/LtJ and ~6x coverage of the other strains**  
**1st data release 1st July**



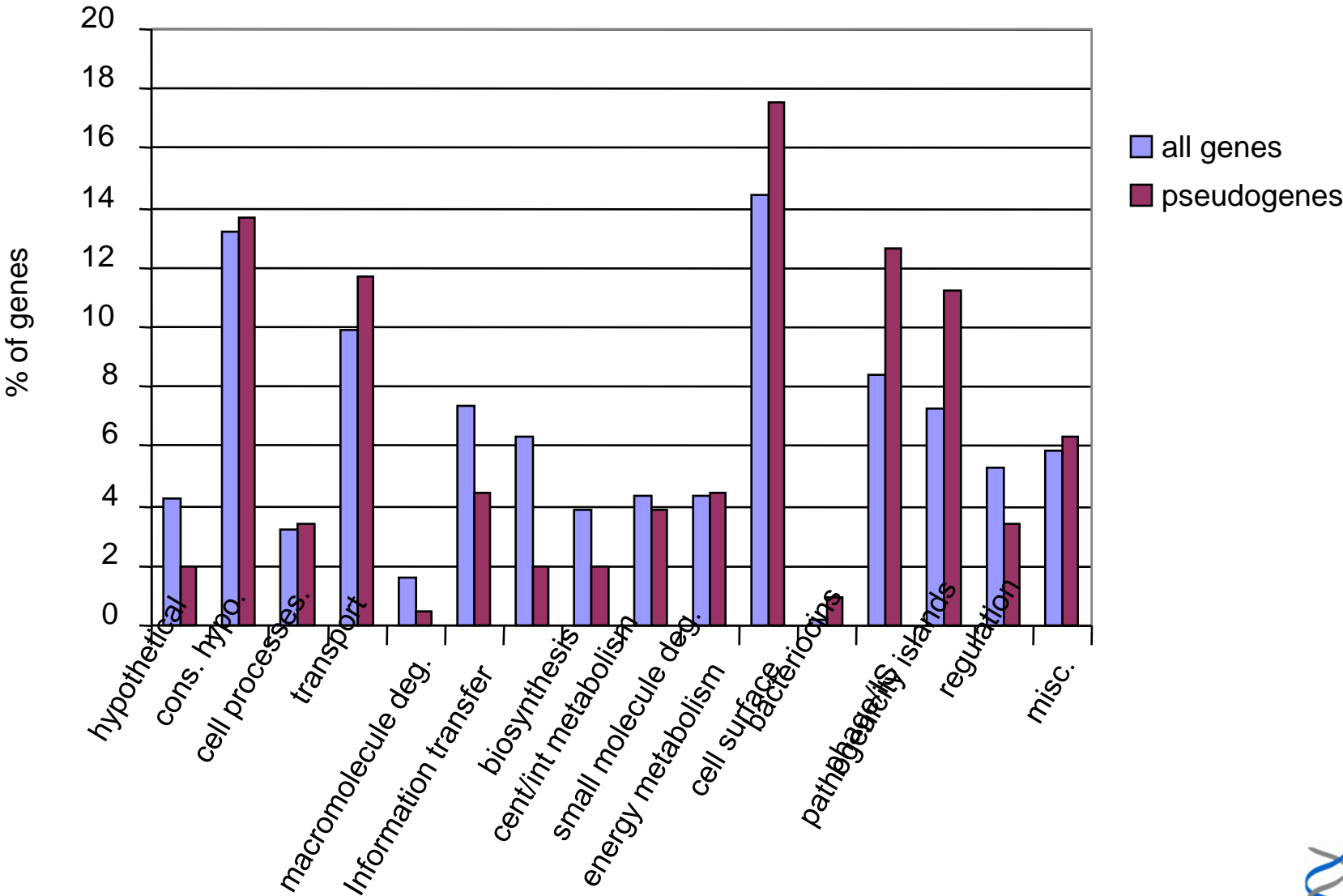
# Microbial applications - *Salmonella enterica* serovar Typhi

18 months sequencing  
10 months analysis



# Salmonella Typhi pseudogenes

204 pseudogenes: 4.4% of coding capacity



# Salmonella Typhi pseudogenes

61% due to single mutations

10% in IS / phage

37% “metabolic”genes

59% lie in unique regions w.r.t. *E. coli*, compared to 33% of all genes

(8% of *S. typhi* unique genes are pseudogenes compared to 2.7% of shared genes)

22% involved in virulence / host interaction, including:

components of 7/12 chaperone/usher fimbrial systems

flagella methylation *fliB*

type-III secreted effectors *sseJ*  
*sopE2*  
*sopA*

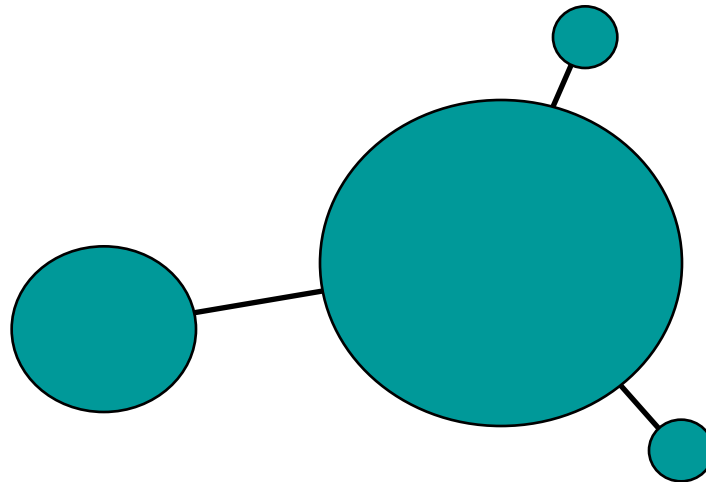
host-range determinants *slrP*  
*shdA/ratA/sivH*

SPI-associated *ttrS* (SPI-2)  
*cigR, marT, misL* (SPI-3)



## Salmonella Typhi phyogeny - MLST

Based on 3 SNPs in 7 genes identified by sequencing





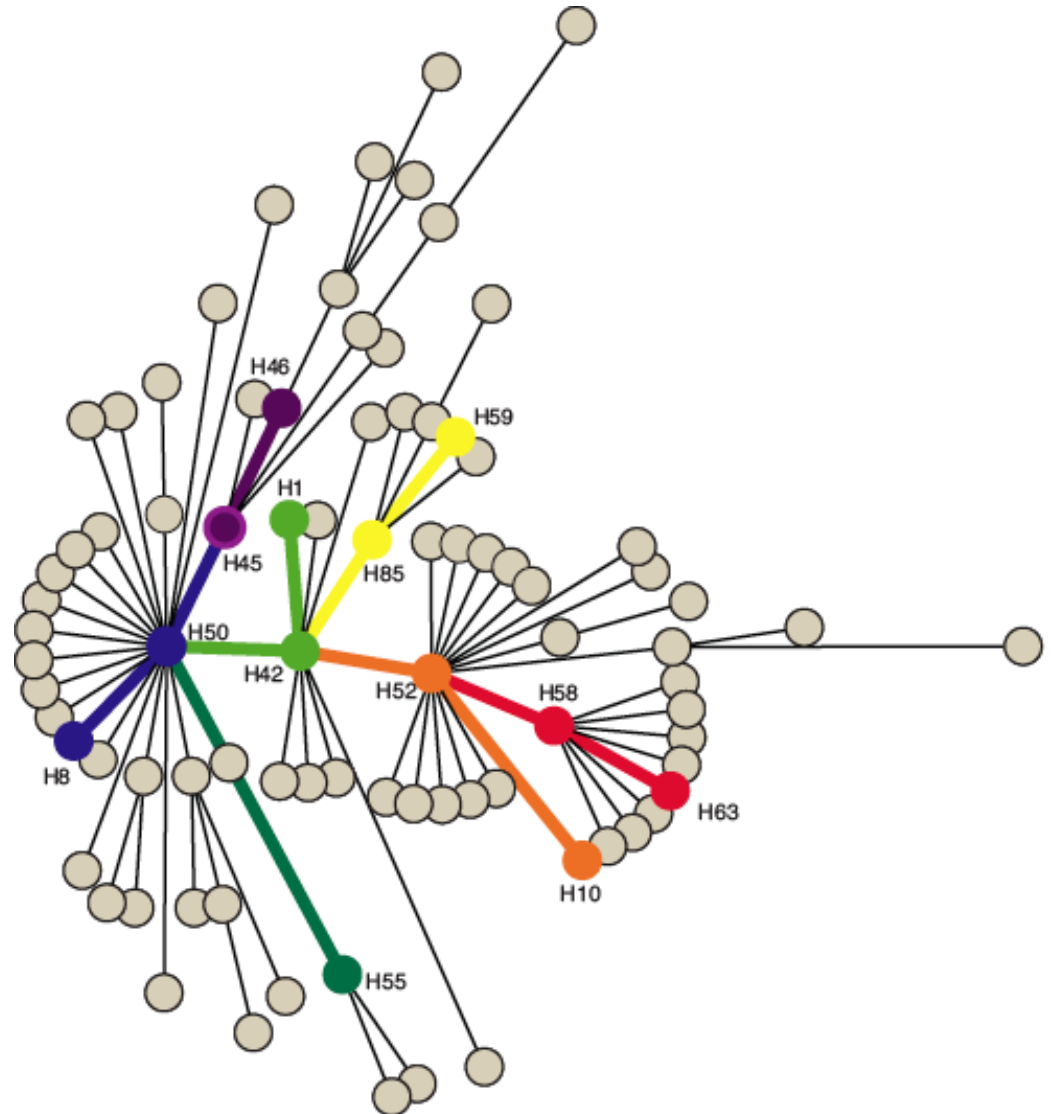
# SNP detection - high throughput re-sequencing

Strains chosen based on 88-SNP tree - major nodes and several H58 strains

Two sequencing platforms:

454/Roche GS20 and GS-FLX

Illumina 1G (Solexa)



# New tech re-sequencing of *Salmonella* Typhi

Strain	Country	Year	Haplotype 454	Solexa	Plasmids
E00-7866	Morocco	2000	H46	10.5 (98.9%)	-
E01-6750	Senegal	2001	H52	8.16 (95.3%)	-
E02-0018	India	2002	H45	13.1 (98.8%)	-
E98-0664	Kenya	1998	H55	10.8 (97.4%)	-
E98-2068	Bangladesh	1998	H42	10.9 (98.4%)	-
J-185SM	Indonesia		H85	13.5 (98.8%)	-
M223		1939	H8	11.1 (99.9%)	-
404ty	Indonesia	1983	H2 (H59)	8.49 (97.0%)	24.6x pBSSB 1a
AG3	Vietnam	2004	H58	10.1 (99.0%)	13.1x -
E98-3139	Mexico	1998	H50	11.1 (95.8%)	5.40x -
150(98)S	Vietnam	1998	H63	-	8.60x -
8(04)N	Vietnam	2004	H58	-	13.1x -
CT18	Vietnam	1993	H1	-	9.80x pHCM1, pHCM2
E02-2759	India	2002	H58	-	65.5x pHCM2
E03-4983	Indonesia	2003	H59	-	7.42x pBSSB 1a
E03-9804	Nepal	2003	H58	-	8.19x pAKU1
ISP-03-07467	Morocco	2003	H58	-	7.87x pAKU1
ISP-04-06979	Central Africa	2004	H58	-	72.9x pAKU1
Ty2-SI	Russia	1916	H10	-	8.60x -
Ty2	Russia	1916	H10	-	-

\*=FLX

Kathryn Holt

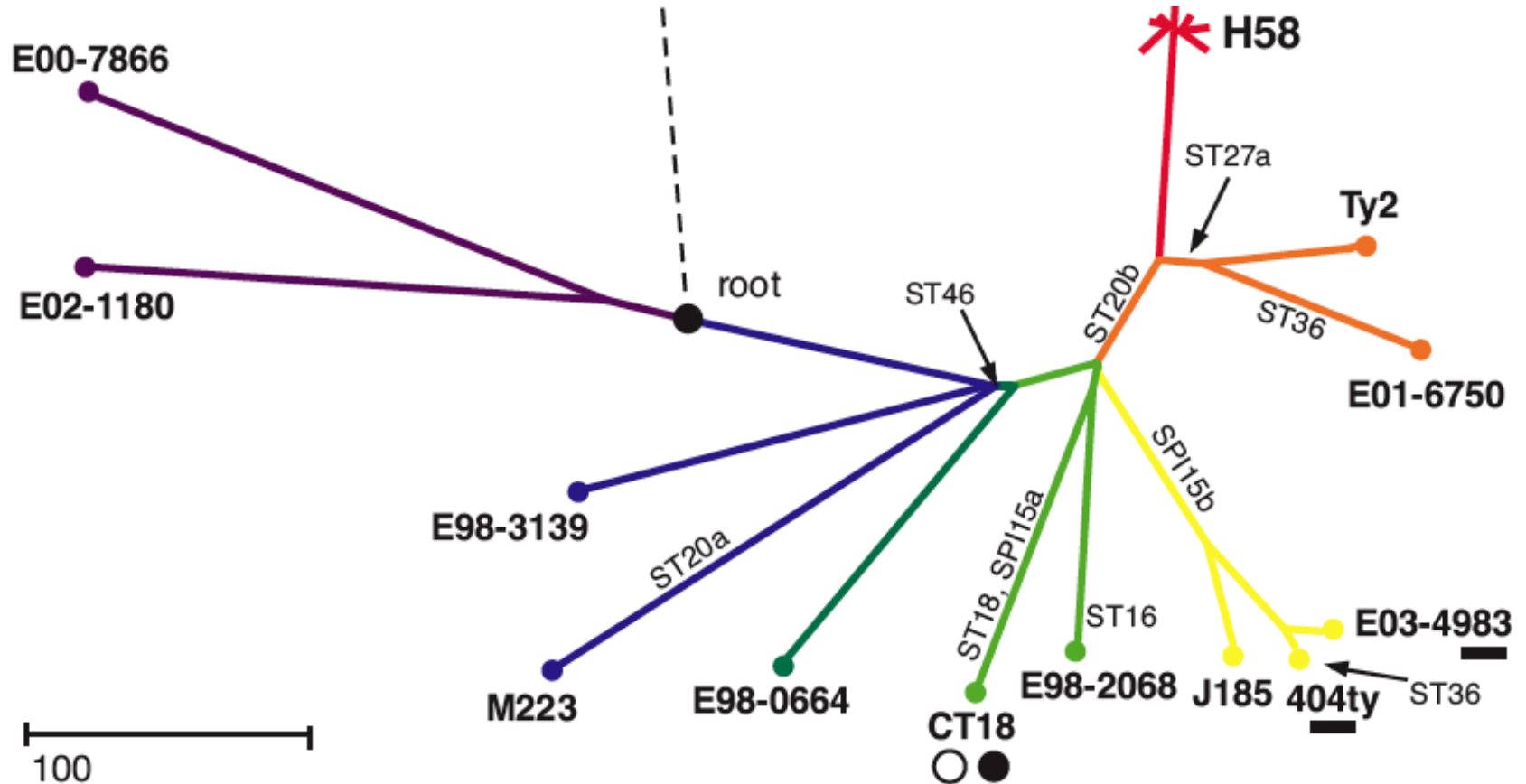


# SNP analysis

1964 SNPs identified after excluding SNPs in repeats and recombined regions.  
1787 SNPs where data was available in all strains.

Only 10 SNPs fail to map to previous tree

- little evidence of recombination within tree, some imports from outside Typhi
- some homoplasy due to selection (2 SNPs in *gyrA*)

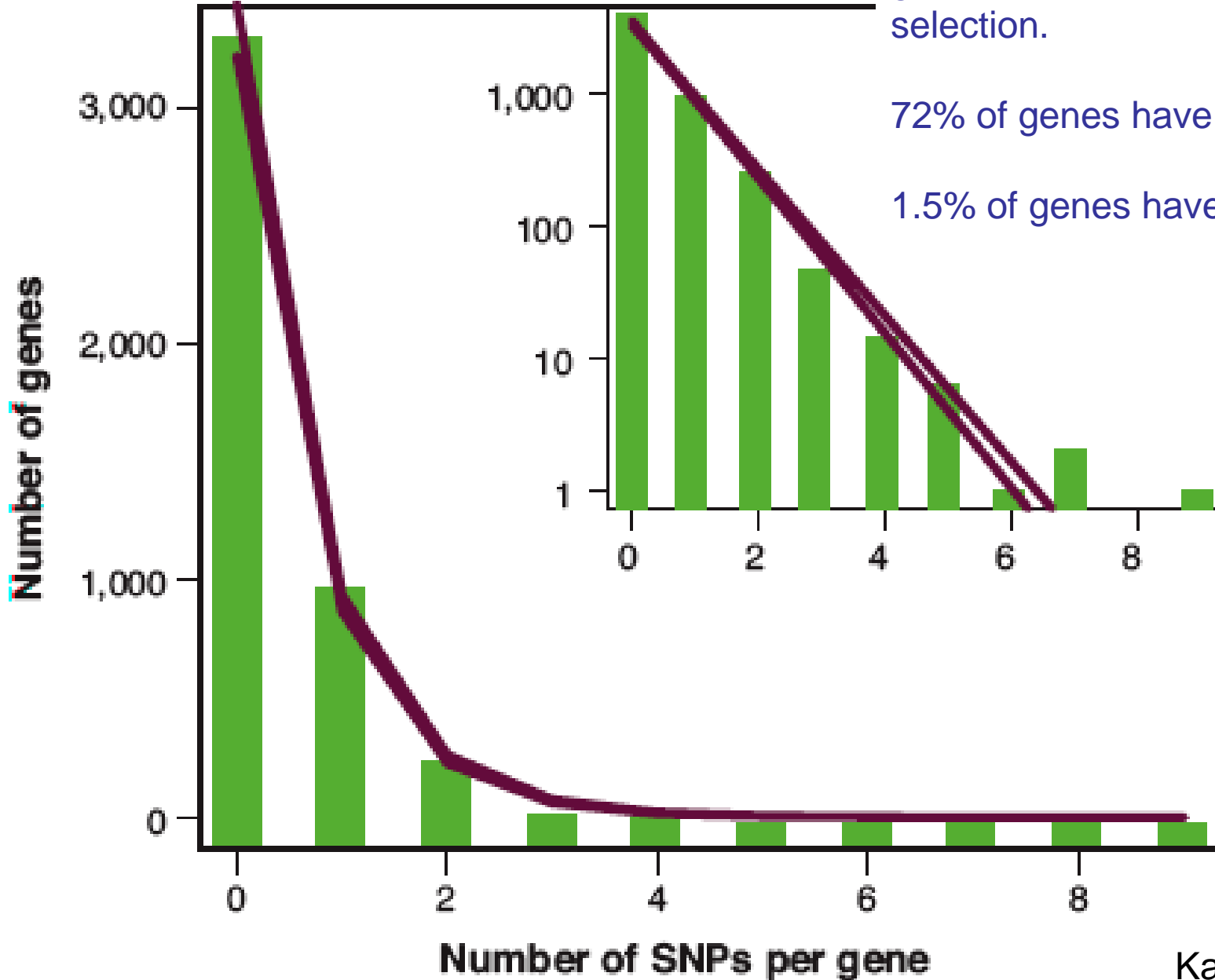


# SNP analysis

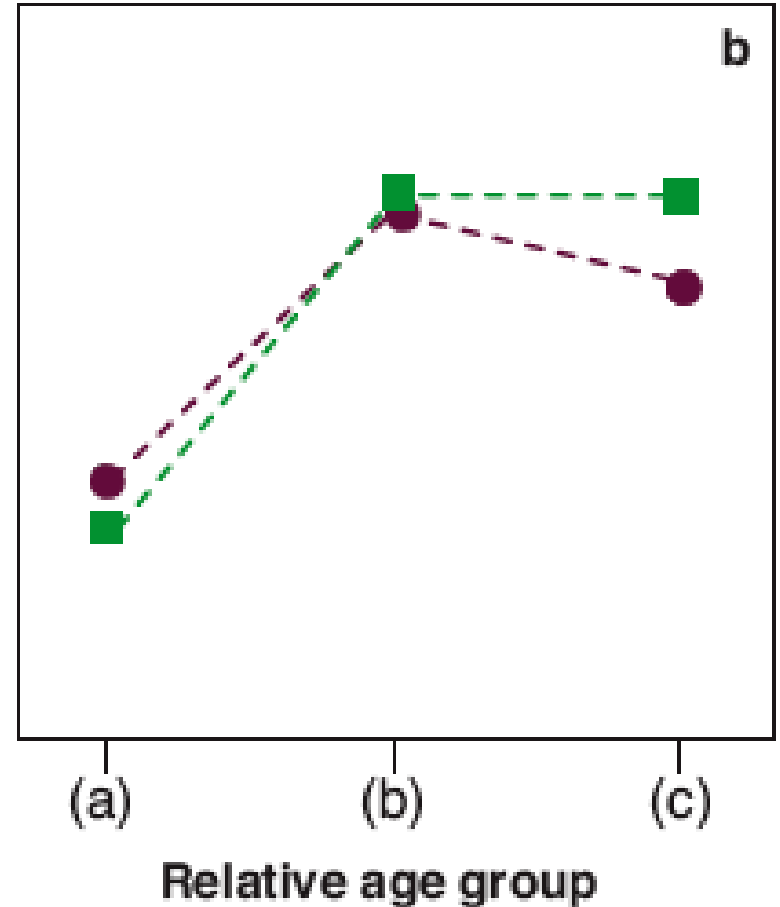
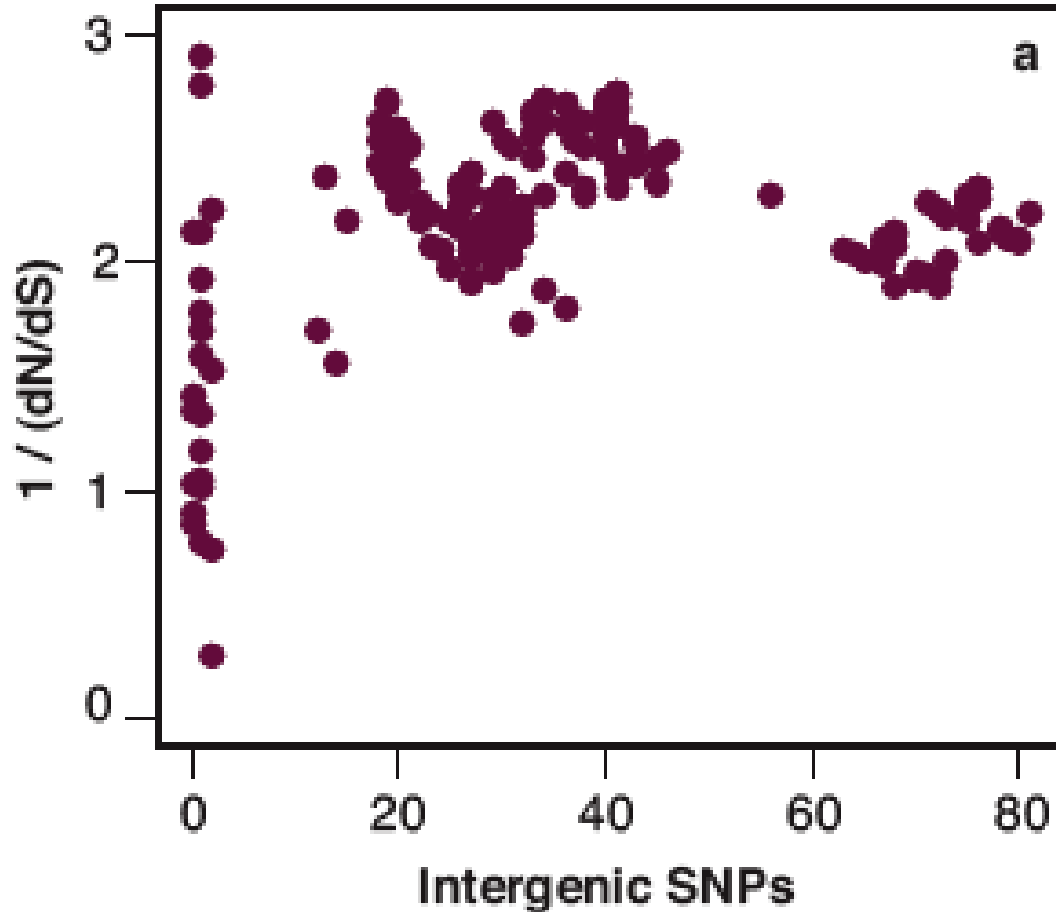
Overall  $dN/dS = 0.476$ : most genes under weak purifying selection.

72% of genes have no SNPs

1.5% of genes have >3 SNPs



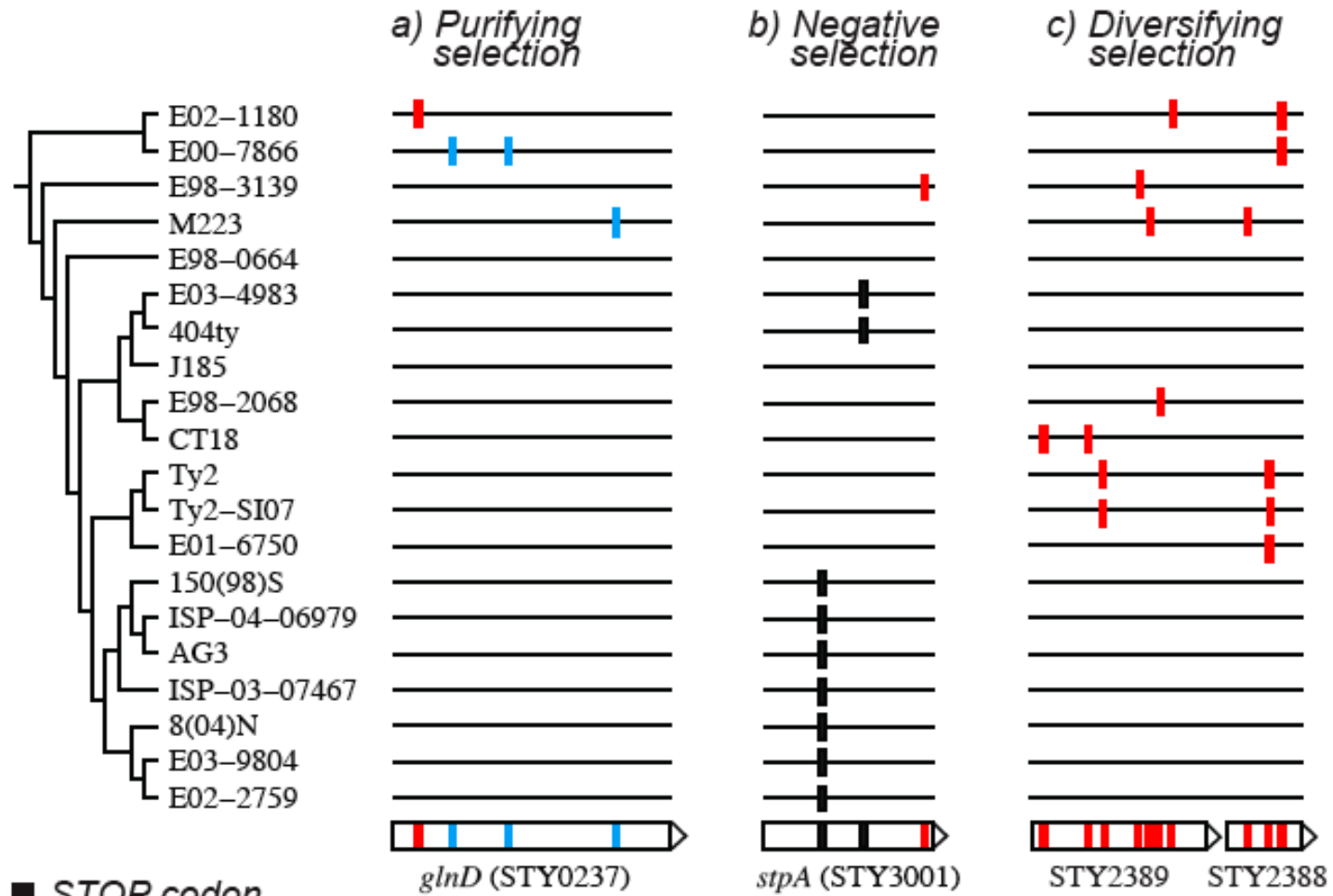
# SNP analysis



Purple: complete allele data  
Green: incomplete allele data



# SNP analysis: genes under selection



Very few examples of potential positive selection



# SNP analysis: genes under selection

Only 26 genes with any evidence for positive selection

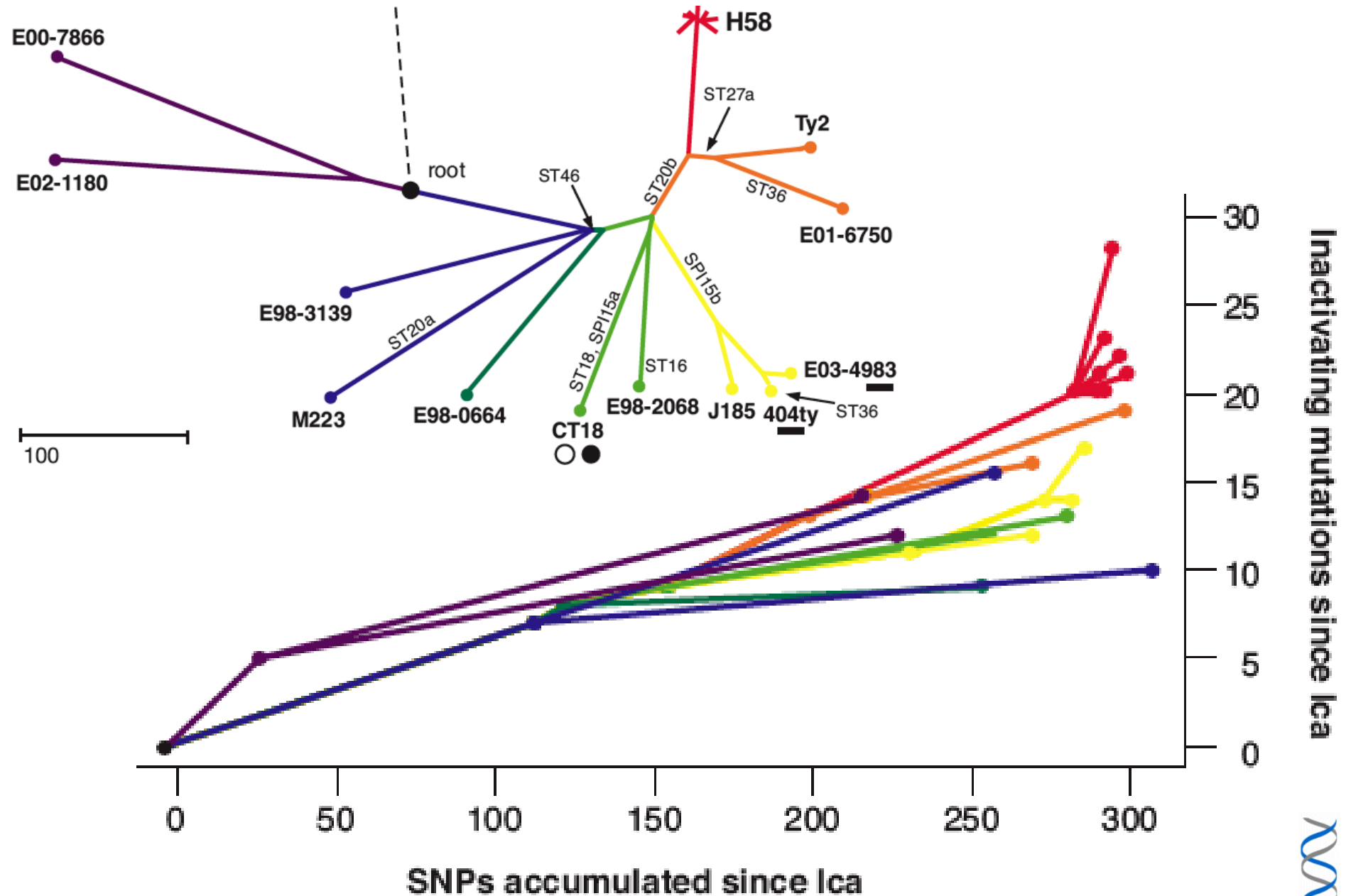
Gene ID	Number of SNPs in coding sequence	nsSNP cluster (residues)	Homoplasic SNPs class (residue)	Gene name	Gene length (residues)	Gene product
STY2389	9*	465,470^	-	<i>yehU</i>	562	two-component system (sensor kinase)
STY2875	7*	-	-		3625	large repetitive protein
STY4656	7*	263,266	-	<i>tviE</i>	579	Vi biosynthesis (polymerization)
STY4318	6*	-	-	<i>bigA</i>	1870	putative surface-exposed virulence protein
STY2499	3	83,87	non x 2	<i>gyrA</i>	879	DNA gyrase subunit A
STY1204	2	-	non (188)		403	putative membrane transporter
STY0194	1	-	non (37)	<i>yadG</i>	309	hypothetical ABC transporter ATP-binding protein
STY0347	1	-	non (563)	<i>tsaC</i>	896	outer membrane fimbrial user protein
STY1689	3	-	syn (35)	<i>ydhD</i>	116	conserved hypothetical protein
STY3775	2	-	syn (418)	<i>priA</i>	733	primosomal protein replication factor
STY1674	1	-	syn (79)	<i>pdxH</i>	219	pyridoxamine 5'-phosphate oxidase
STY3838	0	-	44 bp	<i>fdhD</i>	268	affects formate dehydrogenase-N
STY4805	2	-	186 bp		407	arginine deiminase
STY0042	2	10,11	-		498	putative secreted sulfatase
STY0223	3	47,51	-	<i>hemL</i>	427	glutamate-1-semialdehyde 2,1-aminomutase
STY0565	2	7,8^	-	<i>gcl</i>	594	glyoxylate carboligase
STY0970	3	30,31	-		66	hypothetical protein
STY1264	2	58,59	-	<i>sifA</i>	337	putative virulence determinant
STY1515	2	47,48	-		388	putative oxygenase
STY2388	4	131,131^	-	<i>yehT</i>	240	two-component system (regulator)
STY3222	2	9,12	-		212	possible membrane transport protein
STY3297	2	199,203	-	<i>ordL</i>	434	putative oxidoreductase
STY4161	3	41,44	-	<i>yhjY</i>	235	putative membrane protein
STY4314	5	32,35	-	<i>gph</i>	84	phosphoglycolate phosphatase
STY4890	5	12,12^	-	<i>cstA</i>	717	probable carbon starvation protein (transporter)
STY4659	4	-	-	<i>tviD</i>	832	Vi biosynthesis (polymerization)

\* Deviation from Poisson model of SNPs/gene

^ clustered SNPs in the same strain

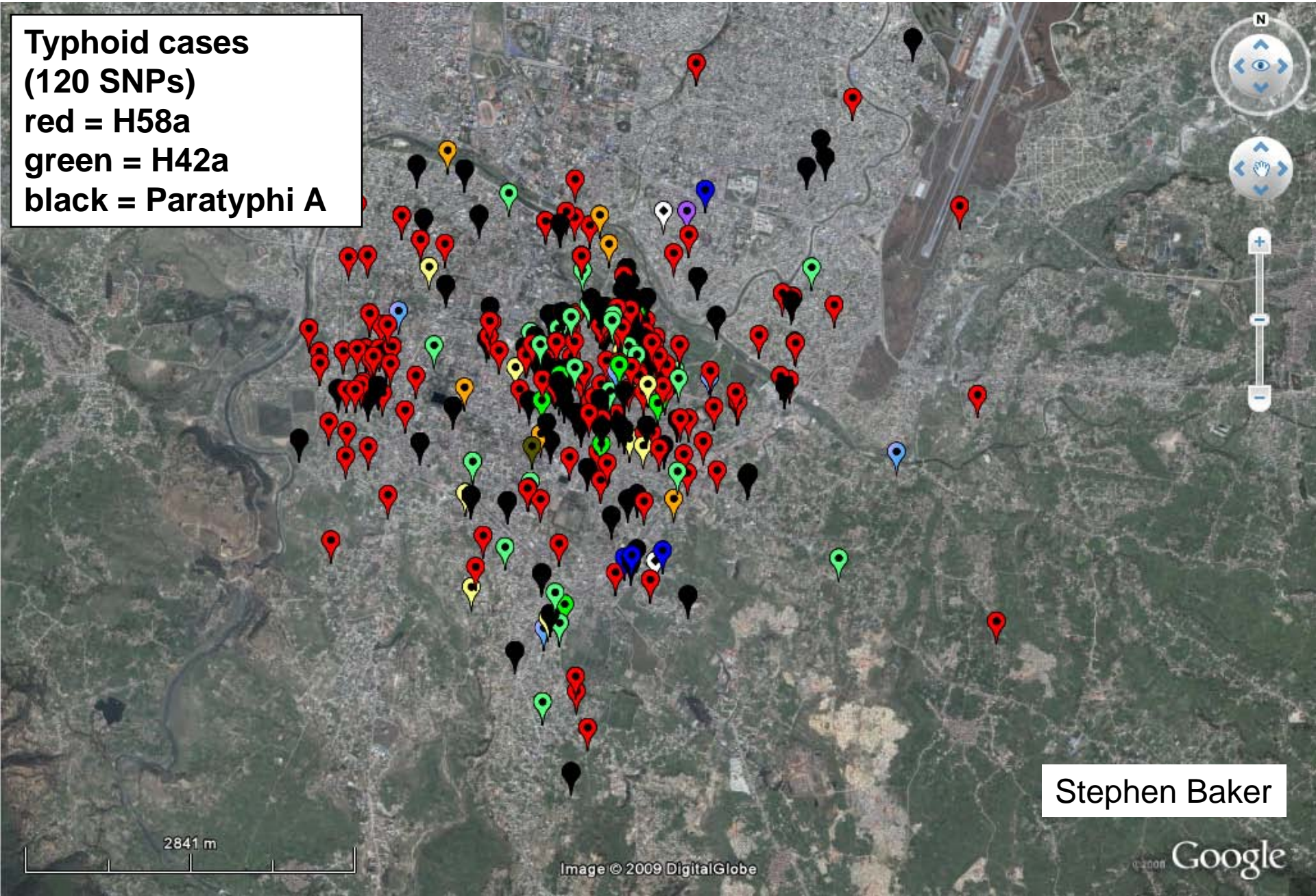


# SNP analysis: New pseudogenes



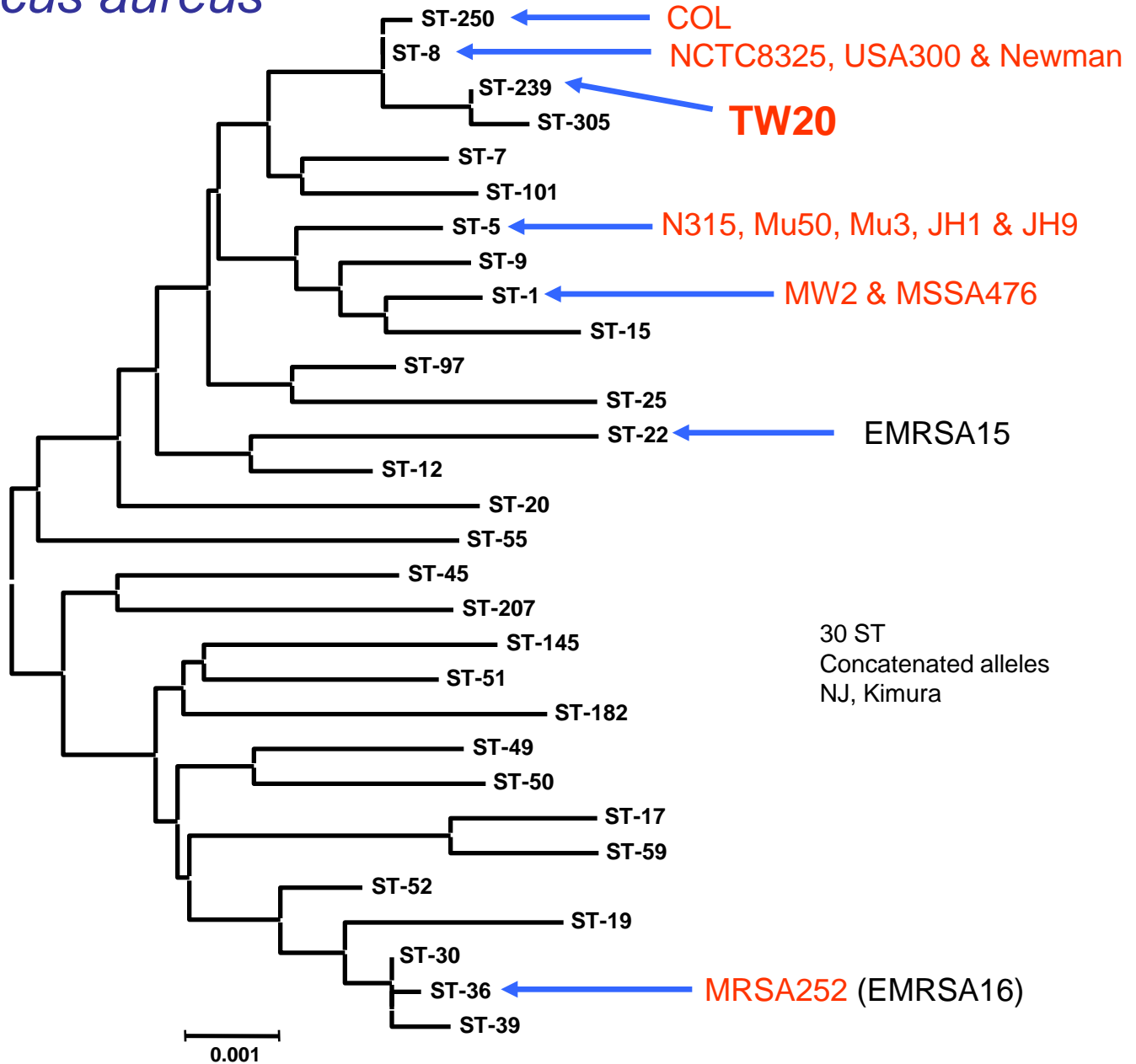
# Putting SNPs to work: epidemiological analysis in the field

**Typhoid cases  
(120 SNPs)**  
red = H58a  
green = H42a  
black = Paratyphi A



Stephen Baker

# Staphylococcus aureus



# Tagged and pooled sequencing of *Staphylococcus aureus* ST239

## Illumina coverage of pools

Combined output from 3 single-end runs

Pool	Total Reads	Reads With Bad Tag	Number of Strains	Predicted Depth	Mapped Depth
1	24,036,478	1,711,455	12	21.77	19.71
2	28,262,372	1,949,455	11	28.00	26.07
3	26,248,675	5,387,374	12	20.35	11.86
4	25,766,633	2,389,248	10	27.36	24.97
5	27,589,770	1,673,168	10	30.33	28.49
6	23,840,904	1,588,954	10	26.04	24.03

Mean depth of mapped regions across all pools = 22.52x



# Tagged and pooled sequencing of *Staphylococcus aureus* ST239

## Illumina coverage of pools

### Pool 1

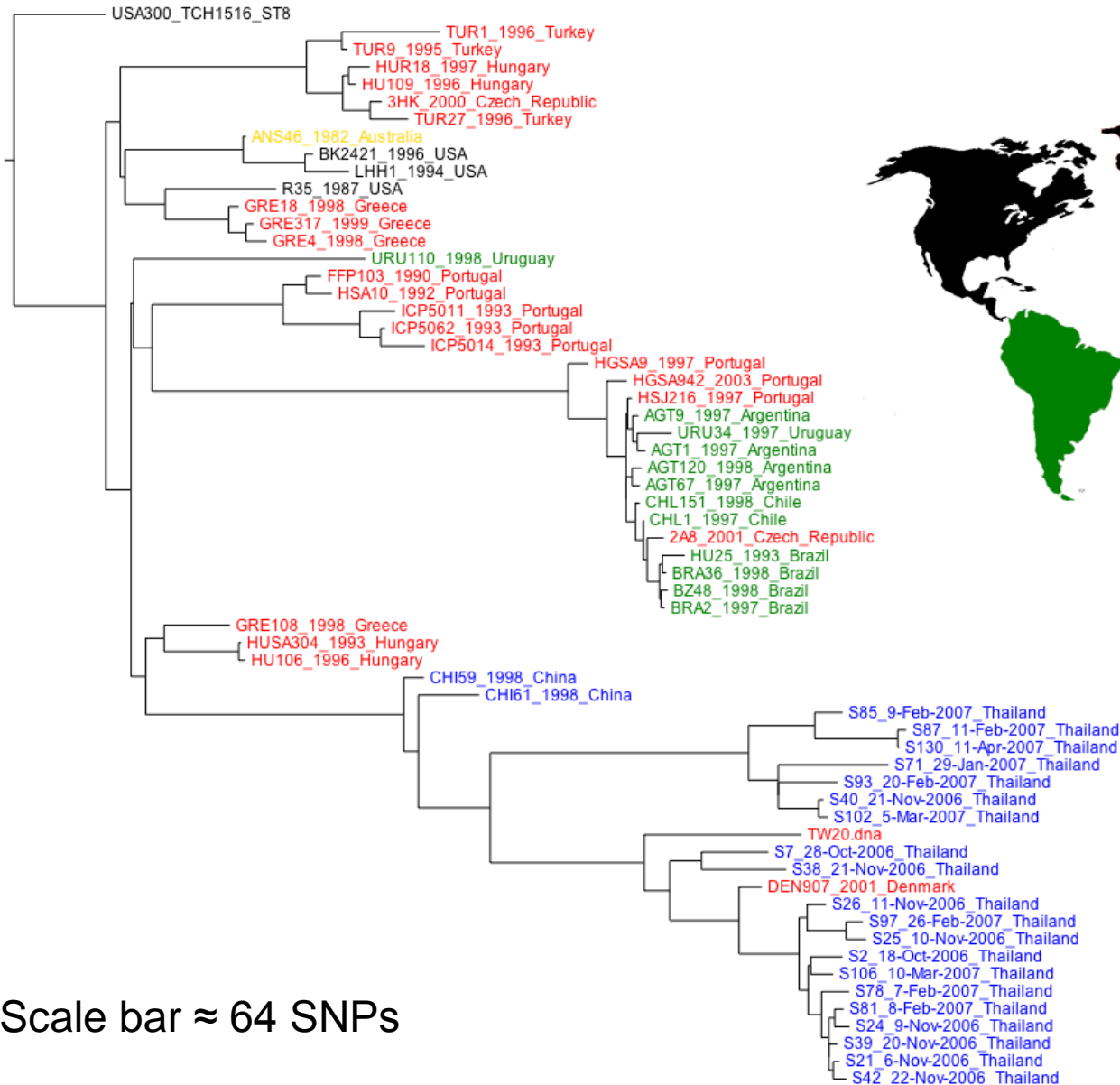
Strain	Reads with tag	Predicted depth	Mapped Reads	% of reference mapped	Mean depth in mapped regions
3HK	2,461,681	28.81	2,016,714	92.96	23.38
BK2491	1,829,408	21.41	1,726,756	94.81	20.09
CHI59	2,245,285	26.28	2,047,861	97.95	23.82
CHI61	2,102,099	24.60	2,016,215	97.60	23.46
FFP103	1,616,205	18.92	1,542,414	91.46	17.95
GRE4	1,677,100	19.63	1,489,725	93.33	17.32
HGSA942	1,964,398	22.99	1,738,857	93.78	20.22
HSJ216	2,076,122	24.30	1,905,849	93.44	22.16
HU109	1,770,961	20.73	1,600,725	93.76	18.60
HU25	1,791,098	20.96	1,713,361	93.92	19.93
ICP5062	1,341,902	15.71	1,305,615	91.68	15.19
TUR9	1,448,764	16.96	1,238,187	93.43	14.39

Note: The reference contains a 120kb insert = 3.9% of length  
This insert is also in the Asian strains (e.g. CHI59 and CHI61)



# Tagged and pooled sequencing of *Staphylococcus aureus*

## Phylogenetic relationships within ST329



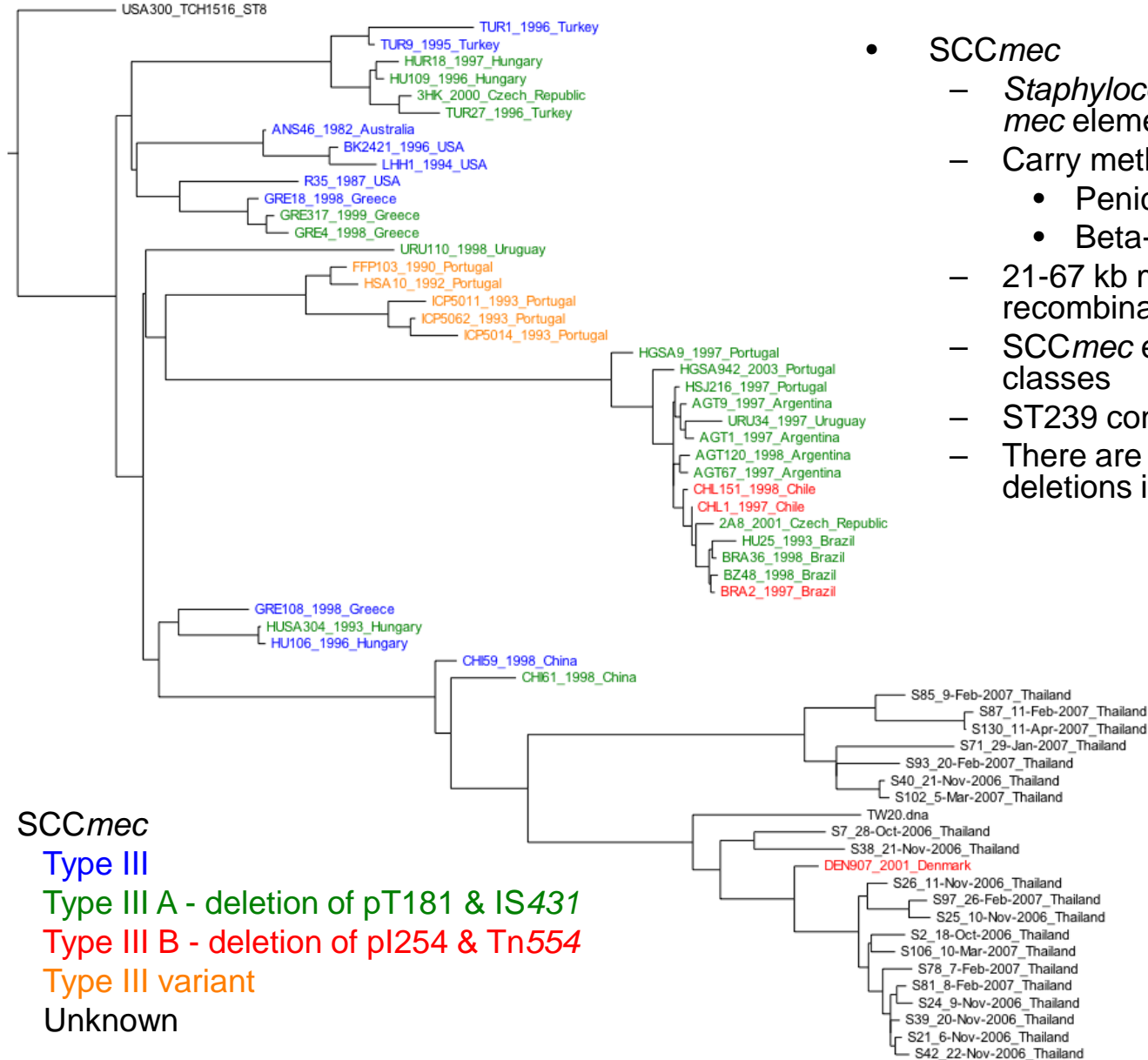
Scale bar  $\approx$  64 SNPs

0.03

Simon Harris



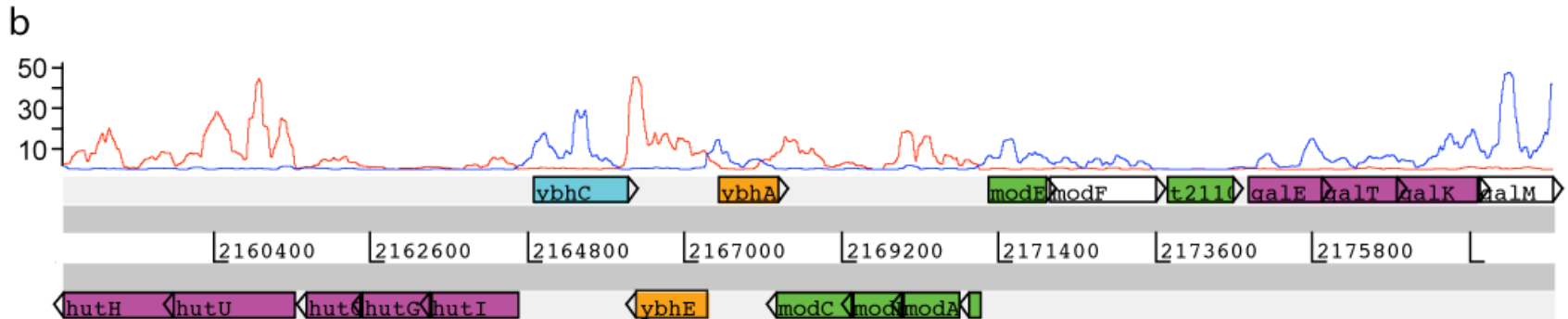
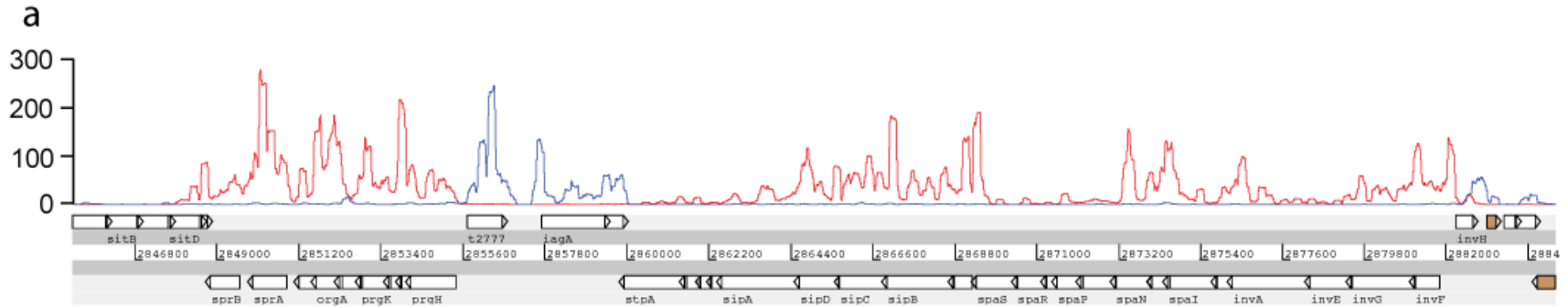
# Tagged and pooled sequencing of *Staphylococcus aureus* Methicillin resistance SCCmec elements



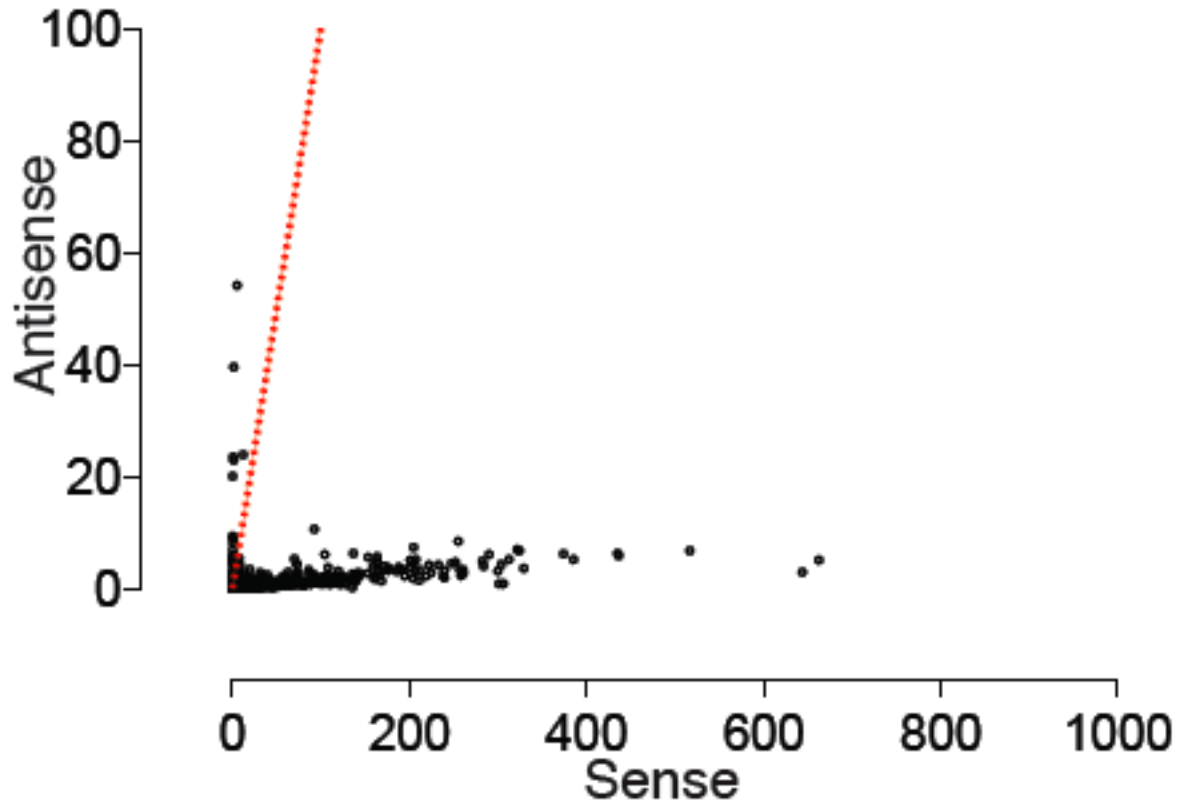
- SCCmec
  - *Staphylococcus* cassette chromosome *mec* element
  - Carry methicillin resistance genes
    - Penicillin binding protein (*mecA*)
    - Beta-lactamase (*blaZ*)
  - 21-67 kb mobile fragment containing recombinase genes
  - SCCmec elements grouped into 7 classes
  - ST239 contains type III
  - There are a number of known deletions in type III SCCmec elements



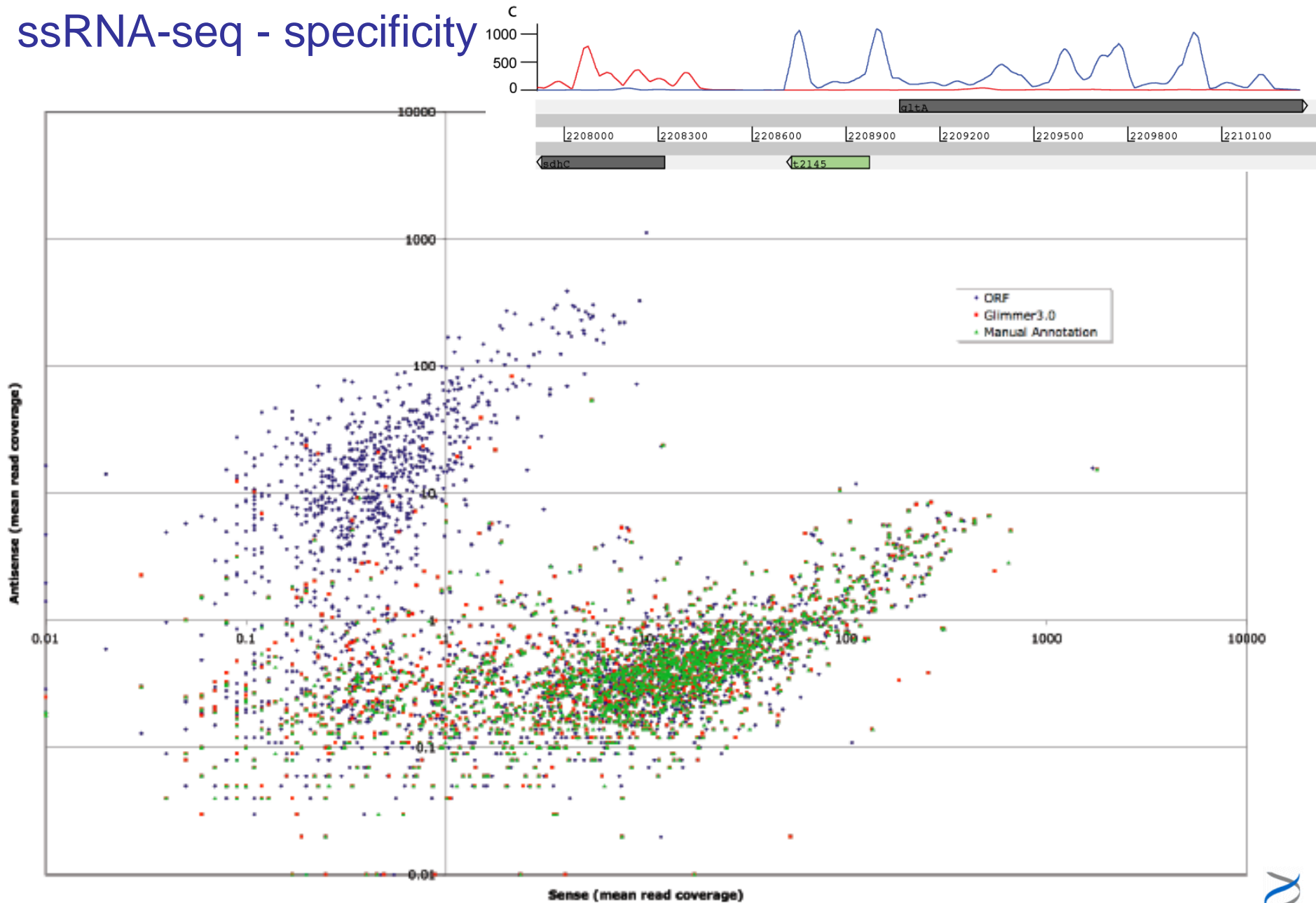
# ssRNA-seq



# ssRNA-seq - specificity



# ssRNA-seq - specificity



Nick Thomson, Tim Perkins, Nick Croucher, Maria Fookes



**WTSI:**

Gordon Dougan

Kathryn Holt

Nick Thomson

Mike Quail

Carol Churcher

Stephen Bentley

Simon Harris

**Max Plank Institute, Berlin/  
University of Cork**

Mark Achtman

Phillipe Roumagnac

**WTSI:**

Nick Croucher

Tim Perkins

Maria Fookes

**OUCRU/ Hospital for  
Tropical Diseases, HCM**

Jeremy Farrar

Christiane Dolecek

Nguyen Tran Chinh

Stephen Baker

**IVI, Seoul**

Camilo Acosta

**INHE, Hanoi**

Thi Anh Hong Le