Chalara fraxinea genomics

Hymenoscyphus pseudoalbidus (anamorph *Chalara fraxinea*, Cf)

evolution,
divergence,
pathogenicity

Mark Blaxter for nornex collaborators meeting February 04-05 2013

The programme has two related themes.

Understanding the pathogenic nature of the fungus

will provide fundamental new information about the pathogen to underpin research on its origins and epidemiology and to inform efforts to develop control measures.

Identifying genetic resistance in ash

will provide molecular markers for identification of genetic resistance in UK trees. These markers are essential for rapid repopulation of devastated areas with resistant stock.

Chalara fraxinea: Planned work

[from proposal]

Genome sequences of up to 30 isolates of the *C. fraxinea* from the UK and Europe, and multiple isolates of the related non-pathogenic fungus *H. albidus*.

Genome sequences will

- reveal the origins of the pathogen,
- provide markers to allow the spread of different strains to be followed
- identify whether the rapid spread is due to crosses with indigenous *H. albidus*
- identify potential virulence factors such as secreted effectors that may be crucial to the invasive nature of the pathogen
- underpin interpretation of transcriptomic data

Comparison of UK isolates to isolates from continental Europe at a whole-genome level will be highly informative.

Complement to existing sequencing efforts by our collaborators in Europe.

Genome sequencing will reveal the origins of the pathogen

While single-marker-, or multiple-marker-based studies can be used to derive estimates of the history of a species, whole genome data, analysed in the context of bayesian approximation or the the coalescent, can reveal

- the history of a population
- the geographical history of a species spread
- the relative age of various splitting events
- the effective size of the population in the past

Genome sequencing will reveal the origins of the pathogen

We will be able to ask

- are UK *C. fraxinea* populations simply a subset of European ones?

- or is there significant population substructure?

- how old is the European *C. fraxinea* epidemic?

- which routes have introduced *C. fraxinea* to Europe and the UK?

- are Asian origin hypotheses supported?

Genome sequencing will provide markers

From comparing the strain genomes we can devise multi-locus genotyping assays that can then be applied to many isolates from across the outbreak range, providing a fine-grained view of the spread and persistence of the fungal genotypes.

- to allow the spread of different strains to be followed
- to assist in mapping loci underpinning traits associated with pathogenesis

Genome sequencing will identify whether rapid spread is due to crosses with indigenous *H. albidus*

We will examine the exciting possibilities that

“pathogenic” loci have introgressed from *H. pseudoalbidus* into *H. albidus*

or that loci underpinning quantitative traits such as survival in different soils or climates have introgressed from resident *H. albidus* into *H. pseudoalbidus*

or that other introgressions (from other taxa) underpin the invasive behaviours observed

Genome sequencing will identify potential virulence factors

We expect that virulence factors will
- evolve faster in the pathogen (to avoid host defences)
- change in expression pattern between non-pathogen and pathogen

These loci can be identified by
- genome comparison of strains, identifying loci under positive selection or subject to selective sweeps
- gene expression that is divergent between pathogenic and nonpathogenic strains

We can also perform prior knowledge-driven searches of the genome data, for example for secreted effectors that may be crucial to the invasive nature of the pathogen

Genome sequencing will underpin interpretation of transcriptomic data

The genome data will permit

- robust inference of transcriptional units through mapping of transcriptome RNASeq data, and

- development of mature annotations (and thus development of systems approaches to understanding *C. fraxinea* biology).

Project data generation plan

"Generate comparative genome sequence data for multiple UK and European isolates."

Location	Number	Lead lab	Sequencing centre
C. fraxinea/H. pseudoalbidus			
UK	1 reference isolate	Kamoun (TSL)	TGAC
UK	~20 isolates	Kamoun (TSL)	TGAC
Europe	~10 isolates	Blaxter (Edinburgh)	GenePool

[additional C. fraxinea genomes being sequenced elsewhere]

Location	Number	Lead lab	Sequencing centre
H. albidus			
UK	5 isolates	Blaxter (Edinburgh)	GenePool

[additional H. albidus genomes being sequenced elsewhere]

Important questions

Which strains to sequence?

some considerations:

We would be well advised to source strains **distinct** in geography and **time** of isolation

For European strains it would be very good to access strains isolated **early** in the epidemic

It would be good to sequence strains for likely **source sites** in Western Europe

Is the sequencing going to be difficult?

No

many strains are already in culture and being grown up.

MALBAC genome amplification methods can be used to increase DNA availability.

Illumina HiSeq and MiSeq instruments can produce vast amounts of high quality data rapidly.

HiSeq2500	HighOutput	1.4 Bn read pairs	100 bases	300 Gb	12 days	**5000** x genome
	RapidRun	160 M read pairs	150 bases	50 Gb	2 days	**800** x genome
MiSeq	Version2	15 M read pairs	250 bases	6 Gb	1.5 days	**100** x genome

Is the analysis going to be difficult?

Complex, yes.

Difficult, no.

Both TGAC and GenePool have access to large compute resources, and skilled staff with experience of assembly and annotation.

Collaborators have extensive skills in fungal genomics, population genomics and genetical genomics/QTL analyses.

French isolates of *Chalara fraxinea*

From Renaud Ioos, Cécile Guinet and Claude Husson
GUINET Cécile <cecile.guinet@anses.fr>
IOOS Renaud <renaud.ioos@anses.fr>
Claude Husson <claude.husson@nancy.inra.fr>

Barcoded using ITS and FG740, Mcm7 and Tsr1 genes (Husson et al. 2011, Eur J Plant Pathol)

Name of isolate	Date of isolation	Origin	Coordinates (N / E)	Host	Collector	Storage location	Identification method	Colleagues
LSVM82	May 2008	Mersuay (France)	47.78 / 6.15	Wood of F. excelsior (stem)	P. Loevenbruck (ANSES, LSV Nancy)	ANSES, LSV Nancy	Morphological features and barcoding	Ioos R. Guinet C., ANSES LSV Nancy
MIG-M-1	April 2009	Migneville (France)	48.53717 / 6.77869	Wood of F. excelsior (stem)	O. Cael (INRA Nancy)	INRA Nancy	Morphological features and barcoding	Marçais B., INRA Nancy
GIR-M-2	April 2009	Girecourt-sur-Durbion (France)	48.26041 / 6.58838	Wood of F. excelsior (stem)	O. Cael (INRA Nancy)	INRA Nancy	Morphological features and barcoding	Marçais B., INRA Nancy
LAN-M-1	April 2009	Languimberg (France)	48.72289 / 6.88468	Wood of F. excelsior (stem)	O. Cael (INRA Nancy)	INRA Nancy	Morphological features and barcoding	Marçais B., INRA Nancy
FON-M-1	April 2009	Fontenoy-le-Chateau (France)	48.00142 / 6.19597	Wood of F. excelsior (stem)	O. Cael (INRA Nancy)	INRA Nancy	Morphological features and barcoding	Marçais B., INRA Nancy

Chalara fraxinea isolate LSVM82

Isolated:	May 2008
Location:	Mersuay (France)
Georeference:	47.78 / 6.15
Source:	Wood of *Fraxinus excelsior* (stem)
Isolated by:	P. Loevenbruck (ANSES, LSV Nancy)

Clonal hyphal culture established and maintained by: ANSES, LSV Nancy
Identification through: Morphological features and ITS barcoding

Cultures bulked up/DNA prepared by: Claude Husson INRA, Ioos R. & Guinet C., ANSES LSV Nancy
Illumina sequencing library prepared by: Anna Montazam, GenePool
MiSeq run: Stewart Laing, GenePool
Bioinformatics analysis: Urmi Trivedi, GenePool; Georgios Koutsovoulos & Ben Elsworth, Blaxter lab

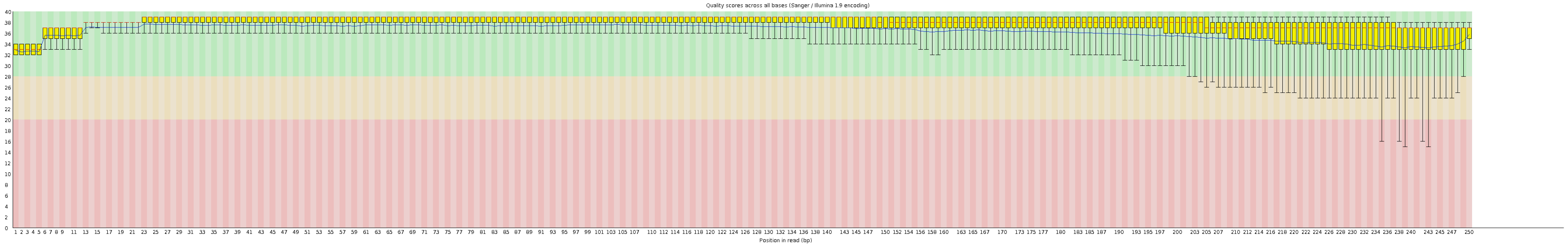
samples arrive and QCd:	Thursday 24th Jan
libraries made and QCd:	Tuesday 29th Jan
MiSeq run:	Wednesday 30th Jan
Data passed QC:	Thursday 31st Jan
Data analysis:	since then...

so what follows is preliminary in the extreme

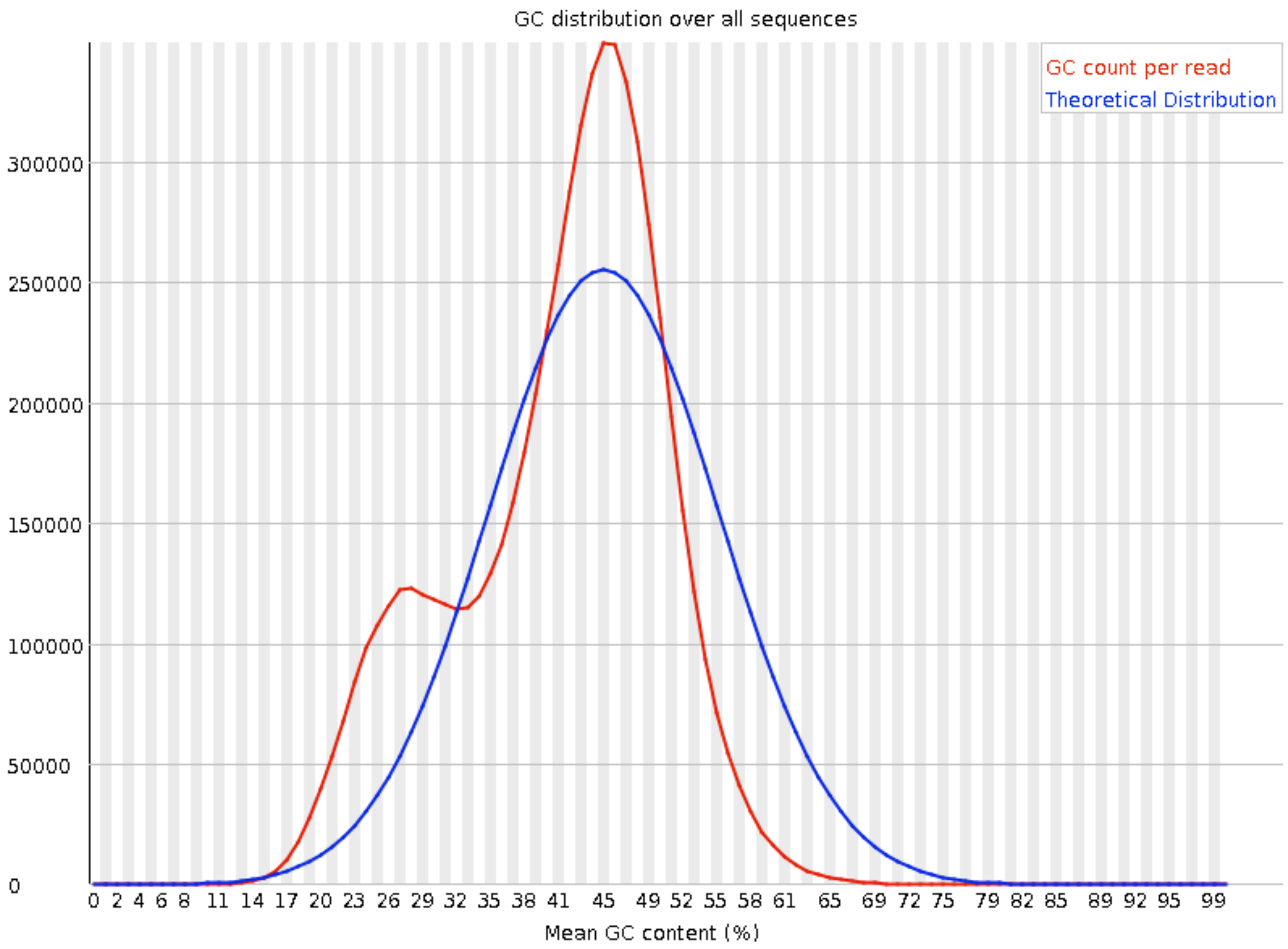
all raw data will be uploaded to ERA/SRA next week
and all analyses and assemblies to the openChalara site asap

Chalara fraxinea isolate LSVM82

Raw data:



25,155,358 reads (12,577,679 pairs) after cleaning
=~ 6 Gbase
=~ 100x coverage

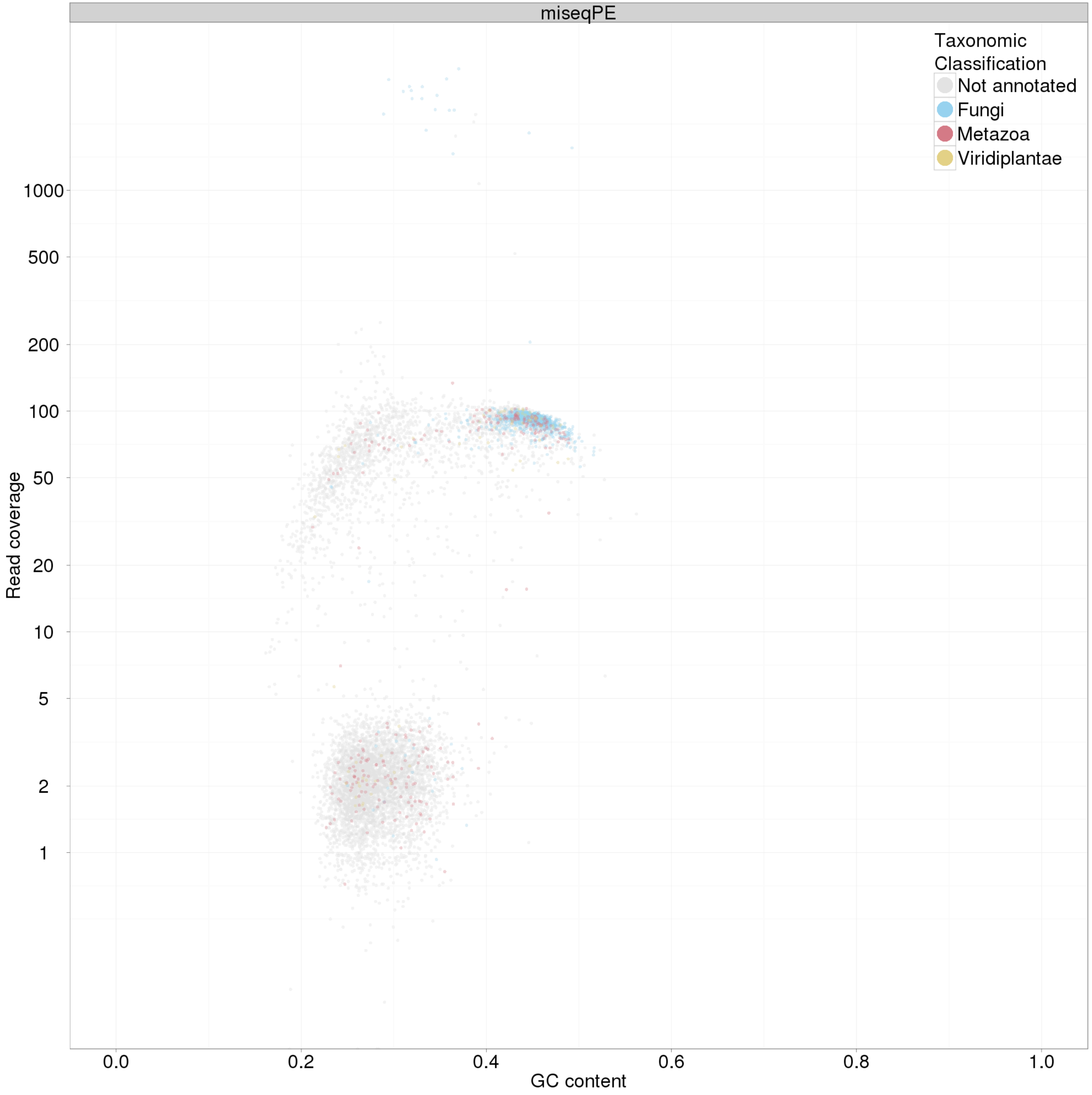


LSVM82

[raw data]

one oddity: double peak in GC content of reads, suggesting contamination.

FASTQC



LSVM82

[raw data]

This "bob plot" shows GC% (x axis) and coverage (y axis) of a preliminary assembly of the data. The genome is (as expected) covered ~100x, but there is a low-level (~3x) "contamination" [easily removed]

github.com/skumar/assemblage

Georgios Koutsovoulos,Edinburgh

Chalara fraxinea isolate LSVM82

mapped (using *smalt*) to the TGAC assembly of the Norfolk KW1 isolate (63 Mb).

Total reads mapped	= 24,830,574	(98.37%)
Duplicate reads	= 455,074	(1.80%)
Reads mapped as proper pairs	= 24,031,770	(95.21%)
Singletons	= 86,074	(0.34%)
Reads with mate mapped to a different contig	= 567,804	(2.24%)

Urmi Trivedi, GenePool

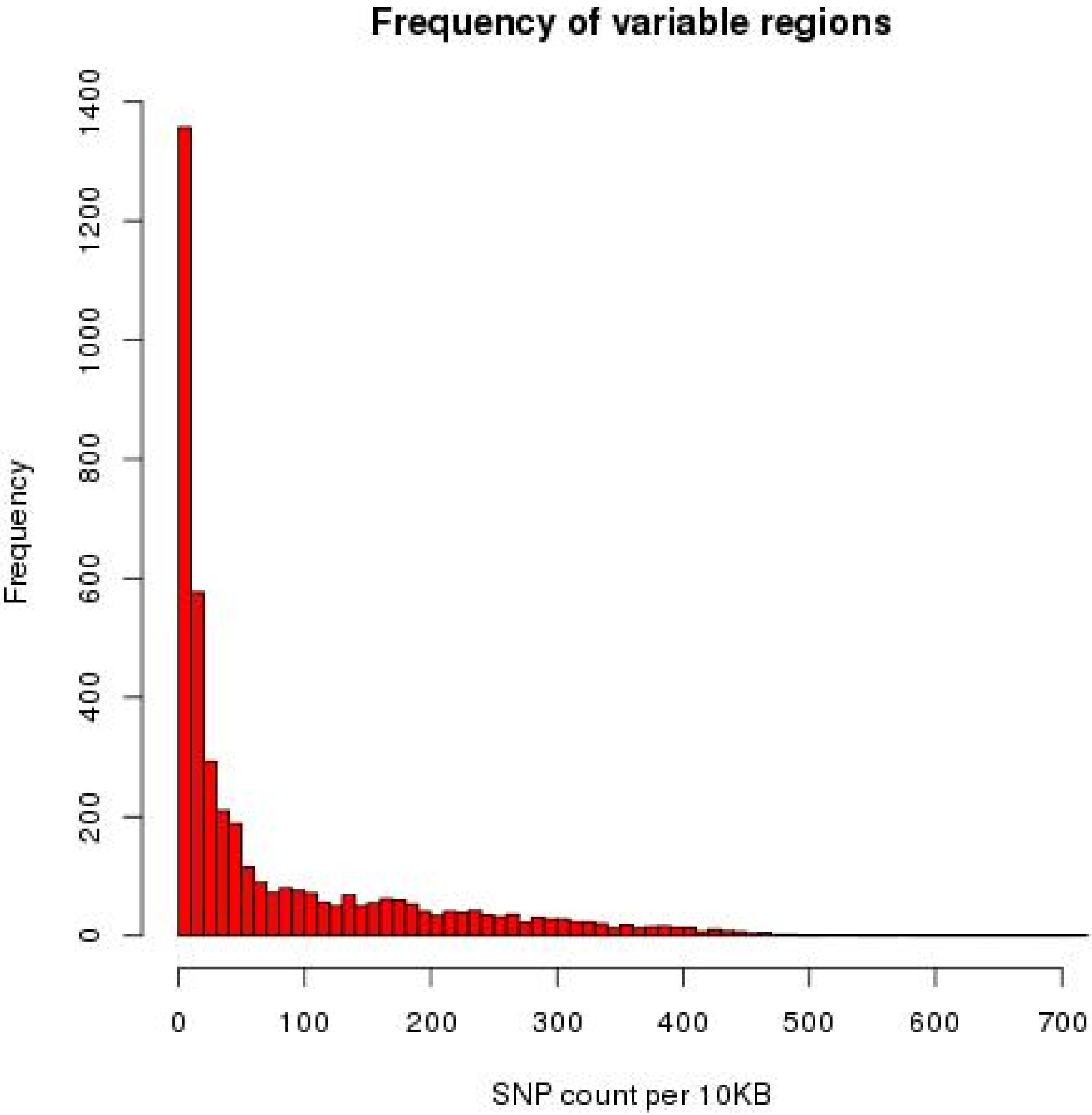
Chalara fraxinea isolate LSVM82

mapped (using *smalt*) to the TGAC assembly of the Norfolk KW1 isolate.

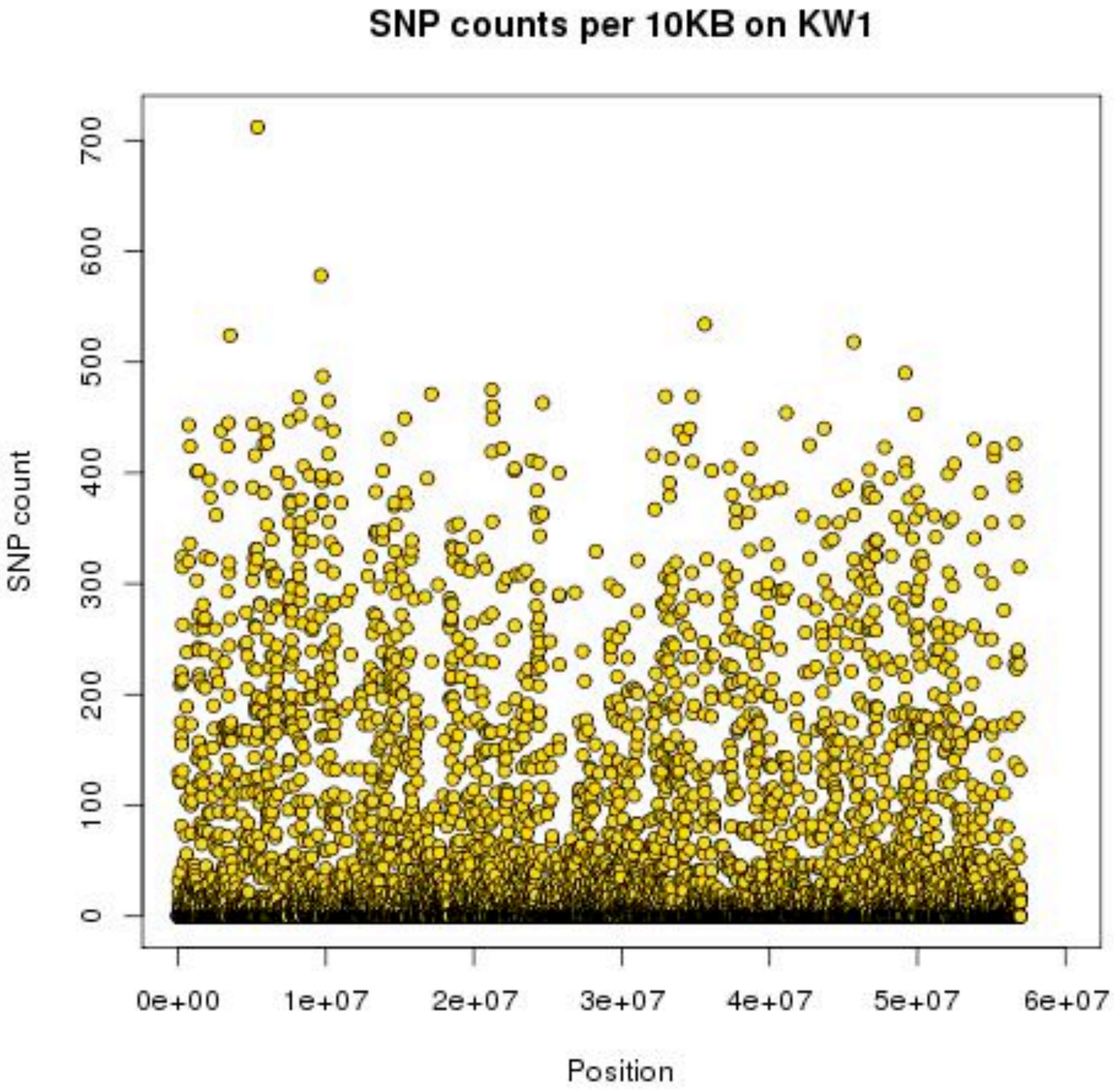
SNP*	316,726	~1 SNP per 200 bases
indels	~11,000°	~2 indels per 10 kb

* after GATK realignment
° (0<x<10 bases)

Urmi Trivedi, GenePool


Frequency of variable regions
Frequency
0
200
400
600
800
1000
1200
1400
0
100
200
300
400
500
600
700
SNP count per 10KB

LSVM82



Mean SNP rate expected to be 2 in 100 bases, or ~40 in 10kb

Many regions have excess SNP mapped (up to 700 per 10 kb).

x axis: KW1 genome (scaffolds >10 kb)

y axis: SNP per 10 kb in nonoverlapping segments

Urmi Trivedi, GenePool

LSVM82 preliminary assembly

Assembly	**C_fraxinea_LSVM82_clc**
contigs (>= 1000 bp)	8,239
Largest contig	203,474
Total length	**58,845,722**
GC (%)	41.80
Reference GC (%)	40.12
N50	32,604
misassemblies	1,472
local misassemblies	1,323
unaligned contigs	4,953 + 406 part
Unaligned contigs length	**7,639,802**
Duplication ratio	1.061
N's per 100 kbp	1,000.7
mismatches per 100 kbp	335.10
indels per 100 kbp	102.98
Largest alignment	192,287
NA50	16,443

Using CLC && paired end information

NOT optimised

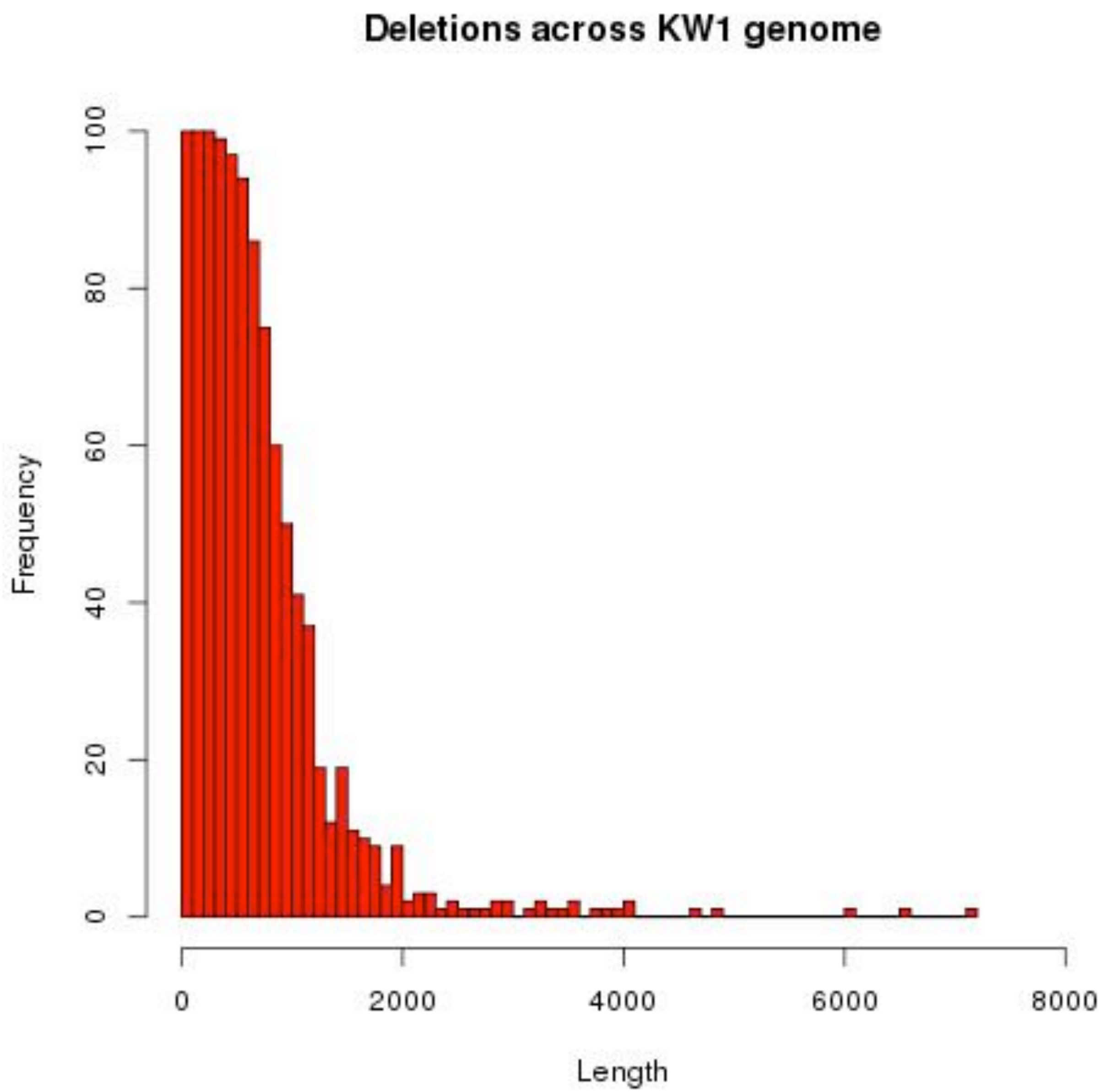
4 Mb smaller than KW1 assembly

Comparison to KW1 reference suggests **7 Mb** additional data*

* NOTE "contaminant" not removed

using QUAST

Georgios Koutsovoulos, Edinburgh



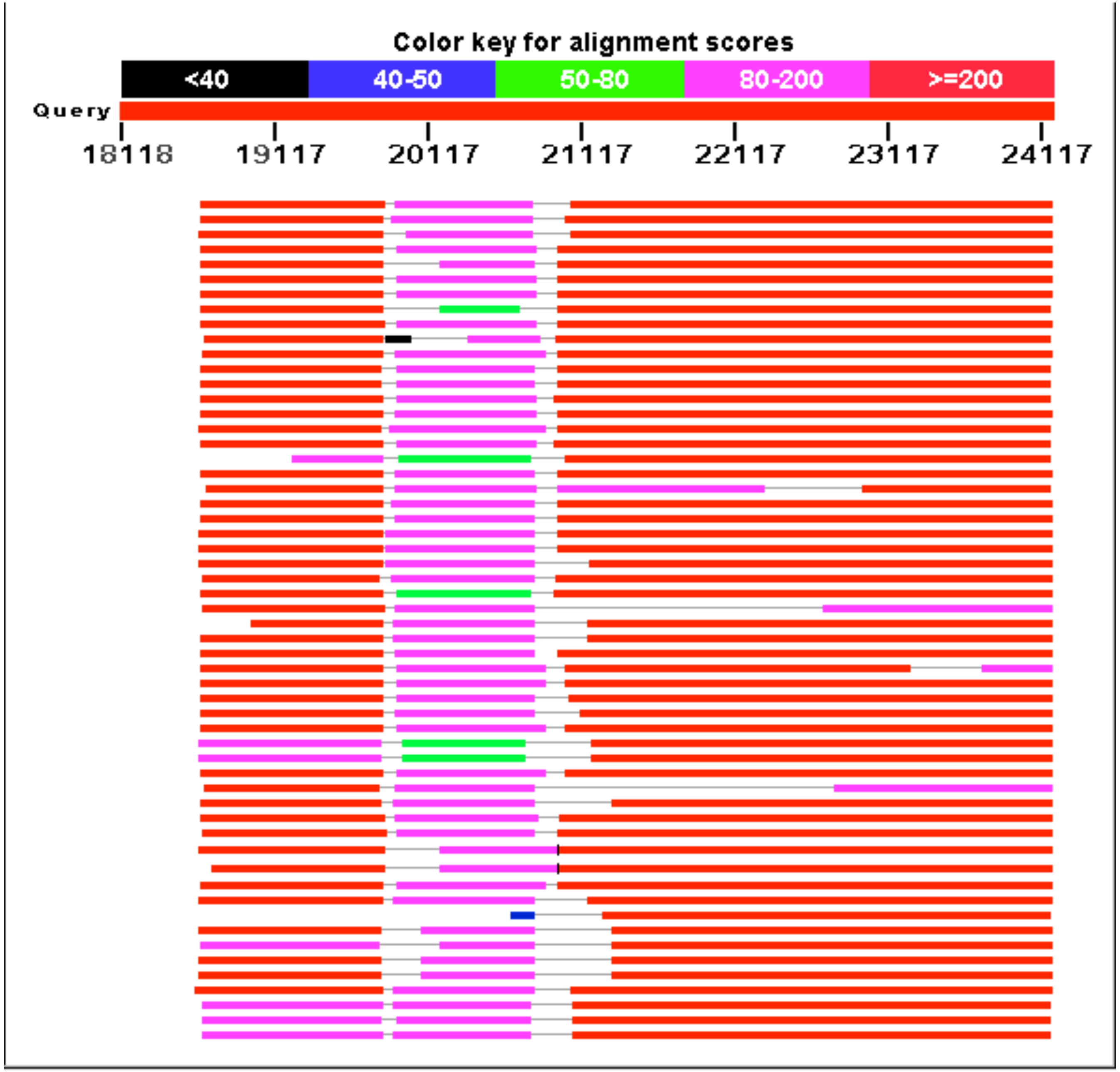
LSVM82

A large number of longer deletions (identified as zero coverage of KW1 scaffolds) are apparent.

x axis: length of zero coverage "deletions"

y axis: frequency of each deletion class

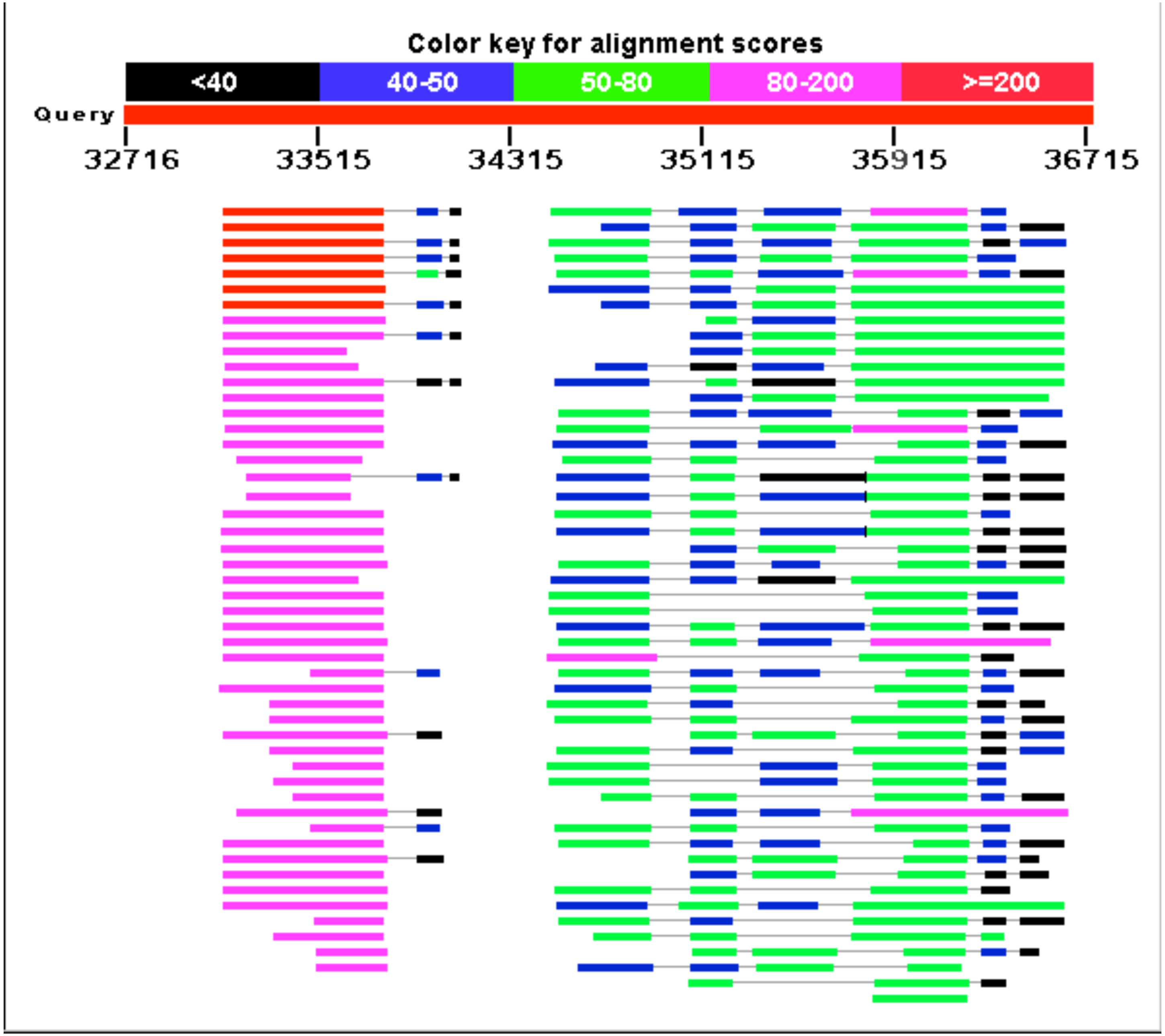
Urmi Trivedi, GenePool



LSVM82

6071 base segment missing compared to KW1

Polyketide synthase



LSVM82

4035 base segment missing compared to KW1

Cytochrome P450

LSVM82 putative deletions span “interesting genes”*

[first 10 deletions...]

KW1 scaffold	start	stop	length	genes in deleted section
Cf746836_TGAC_s1v1_scaffold_1344	9550	16737	7187	Heterokaryon incompatibility domain protein
Cf746836_TGAC_s1v1_scaffold_1306	79579	86113	6534	retroviral POL
Cf746836_TGAC_s1v1_scaffold_737	18118	24189	6071	polyketide synthase (lovastatin nonaketide synthas
Cf746836_TGAC_s1v1_scaffold_752	4140	8957	4817	hypothetical proteins from other fungi
Cf746836_TGAC_s1v1_scaffold_737	27104	31791	4687	hypothetical proteins from other fungi
Cf746836_TGAC_s1v1_scaffold_487	10481	14538	4057	Heterokaryon incompatibility domain protein
Cf746836_TGAC_s1v1_scaffold_737	32716	36751	4035	cytochrome p450 (epsilon hydrolase)
Cf746836_TGAC_s1v1_scaffold_1411	5822	9768	3946	ABC transporter domain
Cf746836_TGAC_s1v1_scaffold_1323	82781	86610	3829	DNA trasposase
Cf746836_TGAC_s1v1_scaffold_752	1	3740	3739	hypothetical proteins from other fungi

Chinese curse: “May you find interesting genes”

Blaxter lab

GenePool (*Karim Gharbi and colleagues*)

TGAC (*Mario Caccamo and colleagues*)

Forest Research (*Steven Hendry and colleagues*)

TSL & others in Nornex

and European colleagues, including

Renaud Ioos, **Claude Husson**