The IWGSC Data repository and wheat data resources hosted at URGI: Overview and perspectives

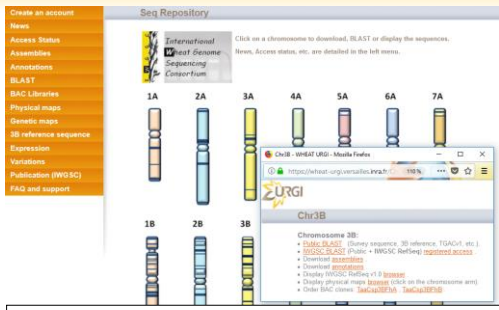
Michael Alaux, Jane Rogers, Thomas Letellier, Raphaël Flores, Cyril Pommier, Nacer Mohellibi, Sophie Durand, Erik Kimmel, Célia Michotey, Mikaël Loaec, et al.

The IWGSC data repository and wheat data resources hosted at URGI: overview and perspectives

Michael Alaux

IWGSC S&P, San Diego, 16 January 2018

IWGSC Data Repository



http://wheat-urgi.versailles.inra.fr/Seq-Repository

Assemblies

- IWGSC RefSeq v1.0 Assembly

• The **IWGSC RefSeq v1.0** (the first version of the reference sequence of the bread wheat variety *Chinese Spring*) assembly is now available for download , BLAST and display in a browser .

The pre-publication data are being made available under the IWGSC General Data Access Agreement which is consistent with the Toronto Agreement and that grants the IWGSC the right to publish the first global analyses of the data. This includes descriptions of whole chromosome or genome-level analyses of genes, gene families, repetitive elements, and comparisons with other organisms.

The **IWGSC RefSeq v1.0 assembly** is an integration of the IWGSC WGA v0.4 – made available in June 2016 – with IWGSC chromosome-based and other resources, including but not limited to:

- Physical maps for all chromosomes;
- Sequenced BACs for 8 chromosomes (1A, 1B, 3B, 3D, 6B, 7A, 7B, 7D) and partial MTP BAC sequences for 2 chromosome arms (4AL, 5BS);
- MTP BAC WGP™ sequence tags for all chromosomes, except 3B;
- BioNano optical maps (7A, 7B, 7DS);
- Alignment to RH maps (D chromosomes); and
- GBS map of the SynOp RIL population CsxRn genetic map (INRA).

With the addition of the resources that have been developed by IWGSC members over the past few years, the quality of the assembly increased substantially. When compared with IWGSC WGA v0.4, the chromosomal scaffold/ superscaffold N50 increased from 7.0 Mb to 22.8 Mb.

The data are available for BLAST searches and can be downloaded .

Assemblies

- Other assemblies available:
  - IWGSC WGA v0.4
  - IWGSC survey sequence (all versions)
  - TGAC v1
  - Other wheat species

Annotations

- IWGSC RefSeq v1.0 Annotation

• **IWGSC RefSeq v1.0 annotation** is available for download and display in a browser .

The pre-publication data are being made available under the IWGSC General Data Access Agreement which is consistent with the Toronto Agreement and that grants the IWGSC the right to publish the first global analyses of the data. This includes descriptions of whole chromosome or genome-level analyses of genes, gene families, repetitive elements, and comparisons with other organisms.

The IWGSC RefSeq v1.0 annotation includes gene models generated by integrating predictions made by INRA-GDEC using Triannot and PGSB using their customised pipeline (previously MIPS pipeline). The integration was undertaken by the Earlham institute (EI), who have also added UTRs to the gene models where supporting data are available. Gene models have been assigned to high confidence (HC) or low confidence (LC) classes based on completeness, similarity to genes represented in protein and DNA databases and repeat content. The automated assignment of functional annotation to genes has been generated by PGSB based on AHRD parameters. In addition, annotated transposable elements (TEs) and non-coding RNAs are available. More information about the annotation data is provided in the README file .

How to access the data?

Access does require registration and agreeing to respect the right of the IWGSC to publish first. For specific access terms, see the IWGSC General Data Access agreement .

- Individuals who have not signed the IWGSC Data Access Agreement should **FIRST** register on the IWGSC website and sign the Agreement. URGI login details will be provided subsequently by email for access to the data. Typically, this it will take no more than 2 business days for your URGI account to be established but occasionally it make take up to a week.
- Individuals who have already signed the IWGSC Data Access Agreement can go directly to the URGI website to access the data using their URGI login details.

Annotations

- Data available:
  - Genes: HighConf and LowConf
  - Functional annotation
  - Transposable elements
  - Markers: ISBP, SNP, DArT, SSR, EST, etc.
  - ncRNAs: miRNA, lncRNA
  - RH maps
  - GBS maps
  - Optical maps

Annotations

- IWGSC RefSeq v1.1 Annotation
  - Refers to the same assembly: the IWGSC RefSeq v1.0 Assembly
  - Data will be available upon publication:
    - Genes: HighConf and LowConf
    - RNA-seq mapping

Access to the IWGSC RefSeq v1.0 data

- Under Toronto agreement (IWGSC general access agreement)

http://www.wheatgenome.org/Tools-and-Resources

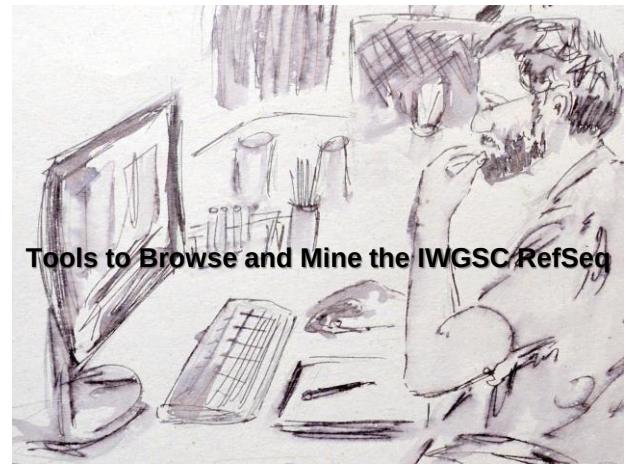
→ will be in open access once published

How to access IWGSC RefSeq v1.0 data?

Access does require registration and agreeing to respect the right of the IWGSC to publish first. For specific access terms, see the IWGSC General Data Access agreement .

- Individuals who have not signed the IWGSC Data Access Agreement should **FIRST** register on the IWGSC website and sign the Agreement; URGI login details will be provided subsequently by email for access to the data. Typically, this it will take no more than 2 business days for your URGI account to be established but occasionally it make take up to a week.

- Individuals who have already signed the IWGSC Data Access Agreement can go directly to the URGI website to access the data using their URGI login details.



Download

IWGSC RefSeq 1.0 assembly

Download

IWGSC RefSeq v1.0 annotation

BLAST

- BLAST dedicated to IWGSC general access agreement:
https://urgi.versailles.inra.fr/blast_iwgsc/?dbgroup=wheat_iwgsc_refseq_v1_chromosomes&program=blastn

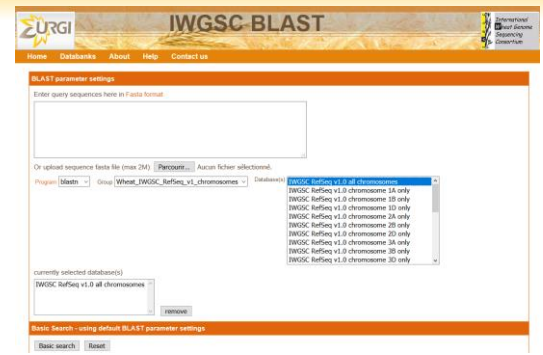
- Allow to BLAST all the available assemblies in one time including the **IWGSC RefSeq v1.0**

- 476k BLAST searches performed in 2017

476k

Michael Alaux

BLAST



https://urgi.versailles.inra.fr/blast_iwgsc/?dbgroup=wheat_iwgsc_refseq_v1_chromosomes&program=blastn

ael Alaux

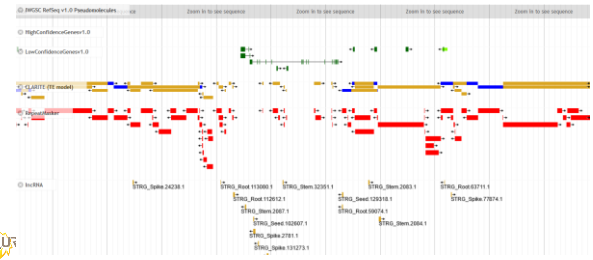
BLAST

- Link to download the matching sequence
- Link to display the JBrowse zoomed in the matching region

Michael Alaux

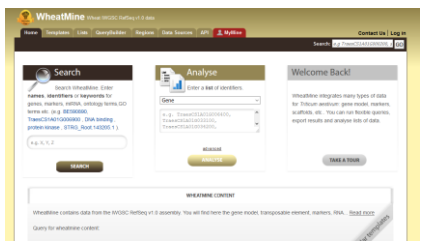
JBrowse

- IWGSC RefSeq v1.0 JBrowse available under general access agreement:
https://urgi.versailles.inra.fr/jbrowseiwgsc/gmod_jbrowse/?data=myData%2FIWGSC_RefSeq_v1.0



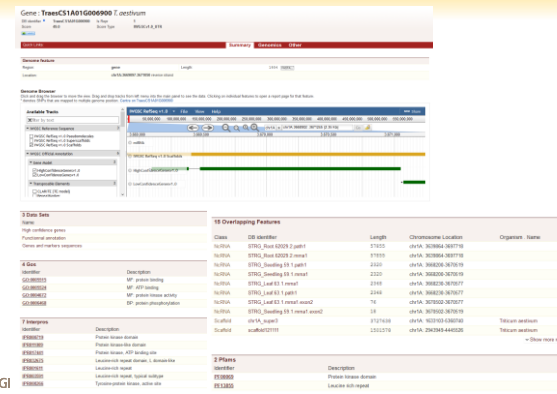
WheatMine

- IWGSC RefSeq v1.0 InterMine available under general access agreement:
https://urgi.versailles.inra.fr/WheatMine



Michael Alaux

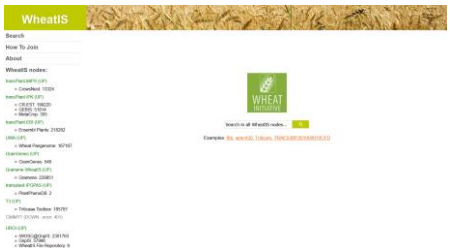
WheatMine



WheatIS search

- IWGSC data can be discovered using the WheatIS search:

  http://www.wheatis.org/Search.php



WheatIS search

- IWGSC RefSeq v1.0 data will be added as soon as they are in open access

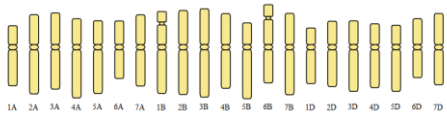
ID	Source	Type	Taxon	Description
JBrowse_chr1A_100233074_100233167_mRNA	IWGSC@GnpIS	SEQUENCE FEATURE	Triticum aestivum	Start = 100233074 [...]
JBrowse_chr1A_101106604_101106809_mRNA	IWGSC@GnpIS	SEQUENCE FEATURE	Triticum aestivum	Start = 101106604 [...]
JBrowse_chr1A_10158696_10158807_mRNA	IWGSC@GnpIS	SEQUENCE FEATURE	Triticum aestivum	Start = 10158696 [...]
JBrowse_chr1A_10231918_10232011_mRNA	IWGSC@GnpIS	SEQUENCE FEATURE	Triticum aestivum	Start = 10231918 [...]
JBrowse_chr1A_102424297_102424426_mRNA	IWGSC@GnpIS	SEQUENCE FEATURE	Triticum aestivum	Start = 102424297 [...]
JBrowse_chr1A_104839251_104839344_mRNA	IWGSC@GnpIS	SEQUENCE FEATURE	Triticum aestivum	Start = 104839251 [...]
JBrowse_chr1A_104839691_104839794_mRNA	IWGSC@GnpIS	SEQUENCE FEATURE	Triticum aestivum	Start = 104839691 [...]
JBrowse_chr1A_105619444_105619537_mRNA	IWGSC@GnpIS	SEQUENCE FEATURE	Triticum aestivum	Start = 105619444 [...]
JBrowse_chr1A_107474939_107475091_mRNA	IWGSC@GnpIS	SEQUENCE FEATURE	Triticum aestivum	Start = 107474939 [...]
JBrowse_chr1A_108071355_108071486_mRNA	IWGSC@GnpIS	SEQUENCE FEATURE	Triticum aestivum	Start = 108071355 [...]

URGI (UP):

- IWGSC@GnpIS: 16385772

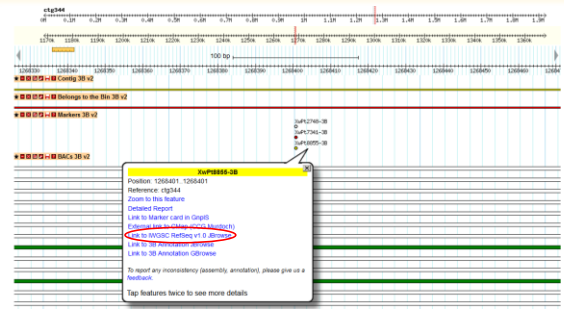
Physical maps

- New versions of 7DS, 1BS added
- All the physical maps are available to display



- Link to the IWGSC RefSeq v1.0

Physical maps



GnpIS-coreDB

- Big data generated from french and european projects

- These data are integrated in the GnpIS-coreDB developed at URGI (part of the GnpIS Information System).

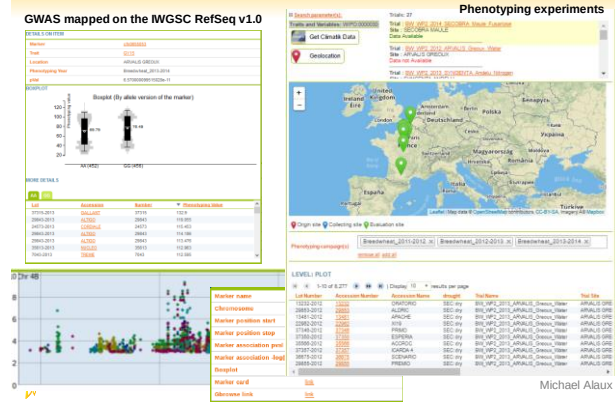
GnpIS-coreDB

- Wheat data overview:

Thematic	Object	#Total	#Open access	#Restricted access to projects
Genetic Resources	Taxon	56	56	0
	Accession	12839	10016	2823
Genetic Maps	Map	30	29	1
	Marker	704822	34164	670658
	QTL	749	465	284
SNP discovery	In Silico Analysis	11	9	2
	Sequence Variation	134904	55362	79542
	SNP, indel	724132	95	724037
Genotyping (high throughput)	Experiment	22	1	21
	Sample	8229	42	8216
	Marker	668543	0	668543
Phenotyping	Trial	853	821	32
	Sample	3660	2985	901
	Variable	291	91	200
		116103		
	Observation	8	527981	633057
GWAS	Analysis	1555	43	1512
	Sample	2365	1839	526
	Variable	359	37	322
	Marker	123866	4109	119757
	Association	824217	48596	775621

http://wheat-urgi.versailles.inra.fr/

Genetic and phenomic data



Integration of the sequence to genetic and phenomic data

- cf. **Alaux *et al.*** companion paper that will be submitted soon.

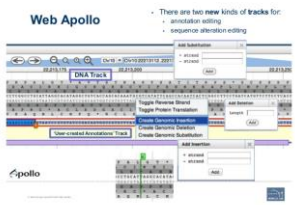
- Use cases:
  - BLAST
  - → Gene (JBrowse)
  - → Marker (GnpIS-coreDB)
  - → QTLs (GnpIS-coreDB)
  - → Phenotyping experiments (GnpIS-coreDB)

Integration of the sequence to genetic and phenomic data

- WheatIS search
- → Gene (WheatMine)
- → Marker (GnpIS-coreDB)
- → GWAS (GnpIS-coreDB)
- → Phenotyping experiments (GnpIS-coreDB)

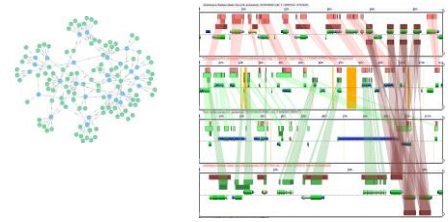
Perspectives

- Management of new/curated versions of the IWGSC RefSeq annotation



Perspectives

- Bioinformatics challenges of a Pan-genome
  - Use new technologies to handle and display large amount of linked data



Acknowledgements

Alaux M.
Letellier T.
Flores R.
Alfama F.
Pommier C.
Mohellibi N.
Durand S.
Kimmel E.
Michotey C.
Guerche C.
Loaec M.
Jamilloux V.
Lainé M.
Adam-Blondon A.F.
Quesneville H.

Choulet F.
Rimbert H.
Leroy P.
Guilhot N.
Paux E.

Rogers J.
Caugant I.
Eversole K.

IWGSC Coordinating Committee

IWGSC Sequencing and Analysis team

All data providers

Questions

IWGSC Data Repository
http://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies

IWGSC BLAST
https://urgi.versailles.inra.fr/blast_iwgsc/?dbgroup=wheat_iwgsc_refseq_v1_chromosomes&program=blastn

Contact me at michael.alaux@inra.fr

@michaelalaux