



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.

► To cite this version:

Michael Alaux, Jane Rogers, Thomas Letellier, Raphaël Flores, Cyril Pommier, et al.. The IWGSC Data repository and wheat data resources hosted at URGI: Overview and perspectives. PAG XXVI - Plant and Animal Genome Conference, Jan 2018, San Diego, United States. pp.32 slides. hal-02788401

HAL Id: hal-02788401

<https://hal.inrae.fr/hal-02788401>

Submitted on 5 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



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

Create an account
News
Access Status
Assemblies
Annotations
BLAST
BAC Libraries
Physical maps
Genetic maps
SB Reference sequence
Expression
Variations
Publication (IWGSC)
FAQ and support

Seq Repository

International Wheat Genome Sequencing Consortium

Click on a chromosome to download, BLAST or display the sequences; News, Access status, etc. are detailed in the left menu.

1A 2A 3A 4A 5A 6A 7A
1B 2B 3B

Chr3B

Chromosome 3B:
• [BLAST](#) (query sequence, 3B reference TGACG1, etc.).
• [Download sequence](#) + IWGSC RefSeq [downloaded access](#)
• [Download assemblies](#)
• [Download annotations](#)
• [Display physical map RefSeq v1.0 browser](#)
• [Display physical maps \(genes tracks on the chromosome arm\)](#)
• [Data file viewer](#): [RefSeqDB](#), [RefSeqDB](#)

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

Michael Alaux

Assemblies

- IWGSC RefSeq v1.0 Assembly

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, 6B, 7A, 7B, 7D) and partial MTP BAC sequences for 2 chromosome arms (5AL, 5BS);
- MTP BAC WGCTM 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 CxsRn 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](#).

[URGI](#) <http://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies> Michael Alaux

Assemblies

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

<http://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies> Michael Alaux

Annotations

- IWGSC RefSeq v1.0 Annotation

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-GDCC using Triannot and PGSS using their customised pipeline (previously MIPS pipeline). The integration was undertaken by the Earlham institute (EJ), 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 PGSS 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 will take no more than 2 business days for your URGi account to be established but occasionally it may 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.

[URGI](#) <http://wheat-urgi.versailles.inra.fr/Seq-Repository/Annotations> Michael Alaux

Michael Alaux

Michael Alaux

Annotations

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

 <http://wheat-urgi.versailles.inra.fr/Seq-Repository/Annotations> Michael Alaux

Michael Alaux

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 [registered on the IwgSC website](#) and sign the Agreement. URGI login details will be provided subsequently by email for access to the data. Typically, this will take no more than 2 business days for your URGI account to be established but occasionally it may 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.

 <http://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies> Michael Alaux

Download

IWGSC RefSeq 1.0 assembly

 https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_RefSeq_Assemblies/v1.0/ Michael Alaux

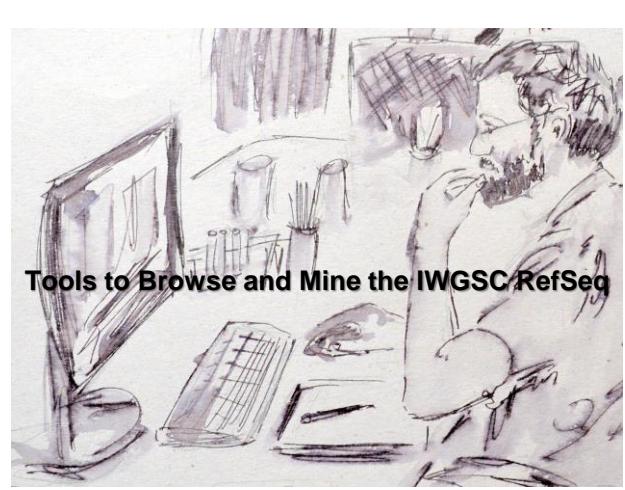
Michael Alaux

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

 URGI <http://wheat-urgi.versailles.inra.fr/Seq-Repository/Annotations>

Michael Alaux



Download

IWGSC RefSeq v1.0 annotation

 https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_RefSeq_Assemblies/v1_0/ Michael Alaux https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_RefSeq_Annotations/v1_0/

Michael Alaux

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



Michael Alaux

BLAST

IWGSC BLAST

The screenshot shows the IWGSC BLAST search interface. At the top, there's a header with the URGI logo and the text "IWGSC BLAST". Below it is a "BLAST parameter settings" section with a text input for "Enter query sequences here in Fasta format". A dropdown menu shows "Program: blastn" and "Database: IWGSC_RefSeq_v1.0 all chromosomes". There's also a "Percom... Aucun fichier sélectionné." button. Below these are "currently selected database(s)" and a "remove" button. At the bottom, there's a "Basic Search - using default BLAST parameter settings" section with "Basic search" and "Reset" buttons.



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

ael Alaux

BLAST

Query	Subject	Score	Identity (Query length)	Percentage	Expect	Start	End
Synt12	IWGSC_RefSeq_v1.0_chromosome_2B_only	2895	15091980 (1980)	99	0.0	1049433	10500000
Synt12	IWGSC_RefSeq_v1.0_chromosome_4A_only	840	630731 (1980)	99	0.0	53966558	53966724
Synt12	IWGSC_RefSeq_v1.0_chromosome_7D_only	545	630732 (1980)	99	0.0	73288459	73299168
Synt12	IWGSC_RefSeq_v1.0_chromosome_5A_only	636	630731 (1980)	99	0.0	651200227	651299533
Synt12	IWGSC_RefSeq_v1.0_chromosome_3D_only	616	630732 (1980)	99	0.0	65094937	65094945
Synt12	IWGSC_RefSeq_v1.0_chromosome_7A_only	814	630744 (1980)	99	0.0	637102978	637142163
Synt12	IWGSC_RefSeq_v1.0_chromosome_2B_only	800	630735 (1980)	99	0.0	2001621	20016108
Synt12	IWGSC_RefSeq_v1.0_chromosome_4D_only	809	630749 (1980)	99	0.0	8736738	87368525
Synt12	IWGSC_RefSeq_v1.0_chromosome_3A-only	305	630730 (1980)	99	0.0	458931002	458930889
Synt12	IWGSC_RefSeq_v1.0_chromosome_5D-only	780	630730 (1980)	99	0.0	541074484	541077178
Synt12	IWGSC_RefSeq_v1.0_chromosome_3D-only	795	630730 (1980)	99	0.0	541335115	541335956
Synt12	IWGSC_RefSeq_v1.0_chromosome_3A-only	795	630730 (1980)	99	0.0	541336983	541336988

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

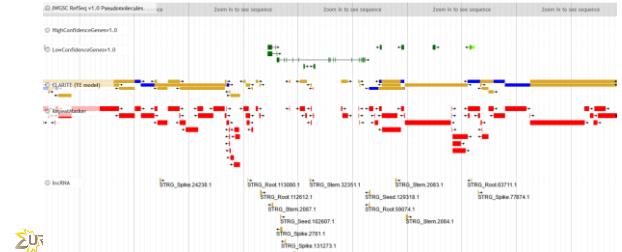


Michael Alaux

JBrowse

- IWGSC RefSeq v1.0 JBrowse available under general access agreement:

https://urgi.versailles.inra.fr/browseliwgsc/gmod_ibrowse/?data=myData%2FIWGSC_RefSeq_v1.0



WheatMine

- IWGSC RefSeq v1.0 InterMine available under general access agreement:

<https://urgi.versailles.inra.fr/WheatMine>

The screenshot shows the WheatMine InterMine interface. It has a search bar at the top with fields for "Search", "Organism", "Gene", "Allele", "Marker", "QTL", and "Variety". Below the search bar are "Analyse" and "Welcome Back!" sections. The "Welcome Back!" section provides an overview of the platform's features: "WheatMine integrates many types of data for wheat varieties: gene model, markers, QTL, transcript, protein, metabolite, and pathway data. You can search, browse, report results and analyse lots of data." At the bottom, it says "WheatMine contains data from the IWGSC RefSeq v1.0 assembly. You will find here the gene models, transposable elements, markers, RNA, ... Take a look! Query for wheatMine content."



Michael Alaux

WheatMine

The screenshot shows a detailed view in the WheatMine interface. At the top, it says "Gene: TriticumAestivumG0000900.T.sorghum". Below it is a "Summary" section with tabs for "Gene", "Allele", "Marker", "QTL", "Transcript", "Protein", "Metabolite", "Pathway", and "Variety". The "Gene" tab is active. It shows a genomic track with various features and a "View Feature" button. Below the track is a "Detailed Features" table with columns for "Name", "Type", "Description", and "Organism". The "Allele" tab is also visible, showing a table with columns for "Allele", "Identifier", "Description", and "Organism". The "Marker" tab shows a table with columns for "Marker", "Identifier", "Description", and "Organism". The "QTL" tab shows a table with columns for "QTL", "Identifier", "Description", and "Organism". The "Transcript" tab shows a table with columns for "Transcript", "Identifier", "Description", and "Organism". The "Protein" tab shows a table with columns for "Protein", "Identifier", "Description", and "Organism". The "Metabolite" tab shows a table with columns for "Metabolite", "Identifier", "Description", and "Organism". The "Pathway" tab shows a table with columns for "Pathway", "Identifier", "Description", and "Organism". The "Variety" tab shows a table with columns for "Variety", "Identifier", "Description", and "Organism".



GnplS-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
SNP discovery	QTL	749	465	284
	In Silico Analysis	11	9	2
Genotyping (high throughput)	Sequence Variation	134904	55362	79542
	SNP, indel	724132	95	724037
Phenotyping	Experiment	22	1	21
	Sample	8229	42	8216
	Marker	668543	0	668543
	Trial	853	821	32
	Sample	3660	2985	901
	Variable	291	91	200
GWAS	Observation	116103	8	527981
	Analysis	1555	43	1512
	Sample	2365	1839	526
	Variable	359	37	322
	Marker	123866	4109	119757
	Association	824217	48596	775621

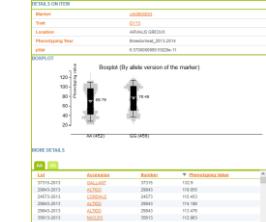


<http://wheat-urgi.versailles.inra.fr/>

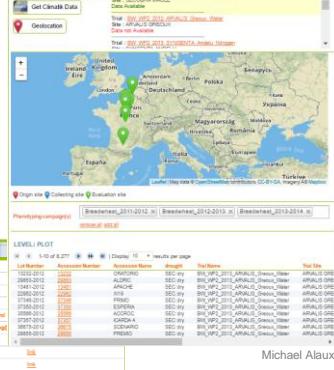
Michael Alaux

Genetic and phenomic data

GWAS mapped on the IWGSC RefSeq v1.0



Phenotyping experiments



Michael Alaux

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 (GnplS-coreDB)
 - QTLs (GnplS-coreDB)
 - Phenotyping experiments (GnplS-coreDB)



Michael Alaux



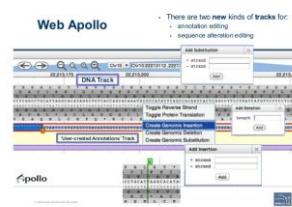
Michael Alaux

Integration of the sequence to genetic and phenomic data

- WheatS search
→ Gene (WheatMine)
→ Marker (GnplS-coreDB)
→ GWAS (GnplS-coreDB)
→ Phenotyping experiments (GnplS-coreDB)

Perspectives

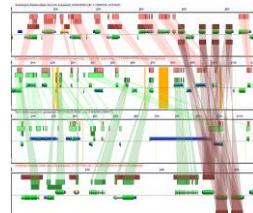
- Management of new/curated versions of the IWGSC RefSeq annotation



Michael Alaux

Perspectives

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



Michael Alaux

Acknowledgements



Alaux M.
Letellier T.
Flores R.
Alfama F.
Pommier C.
Mohellebi N.
Durand S.
Kimmel E.
Michotey C.
Guerche C.
Loaec M.
Jamiloux 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



Michael Alaux



Questions

IWGSC Data Repository

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

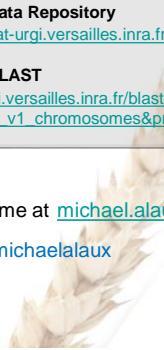
IWGSC BLAST

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

Contact me at michael.alaux@inra.fr



@michaelalaux



Michael Alaux