

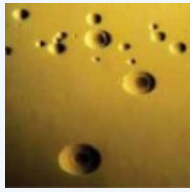


cgfb

BIOINFORMATIQUE

Bacterial genome finishing with Mix and the MolliGen database

« Microbial Bioinformatics day »
Institut Pasteur - 25/09/14
BARRE Aurélien – SEVIN Emeric



Mycoplasma study

- Bacteria lacking a cell wall and parasites for humans, animals, plants or insects
- Small genome → 1 Gb
- Unaffected by many common antibiotics
- Cell culture contamination
- Minimum cell and synthetic biology project based on *M. genitalium* (C. Venter)
- Local collaboration with INRA 1332 - GDPP





MolliGen

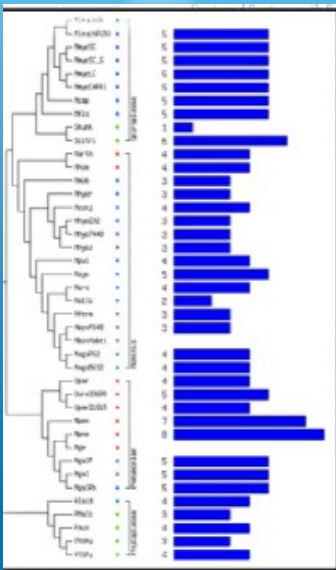
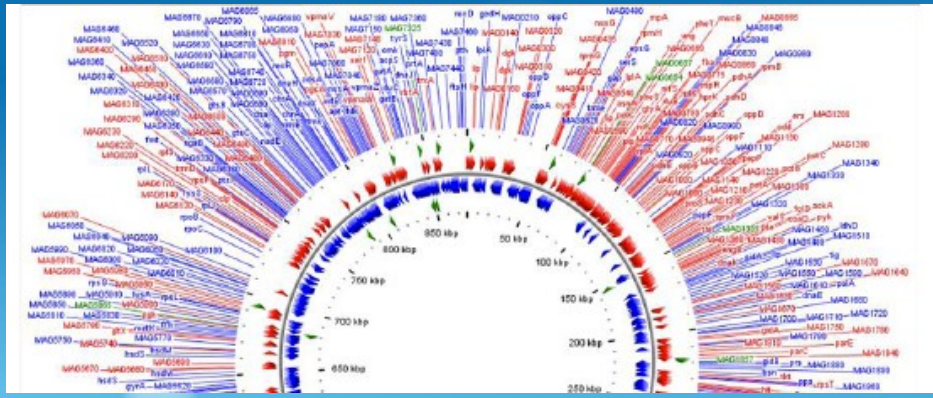
- Web platform dedicated to the comparative genomics of Mollicutes www.molligen.org
- Evolution of the Pasteur « xxxList » web sites
- Over 60 genomes stored, with multiple strains
- Biological experts for curation and annotation
- Large toolbox to analyse and compare genomes
- Private section



cgfb

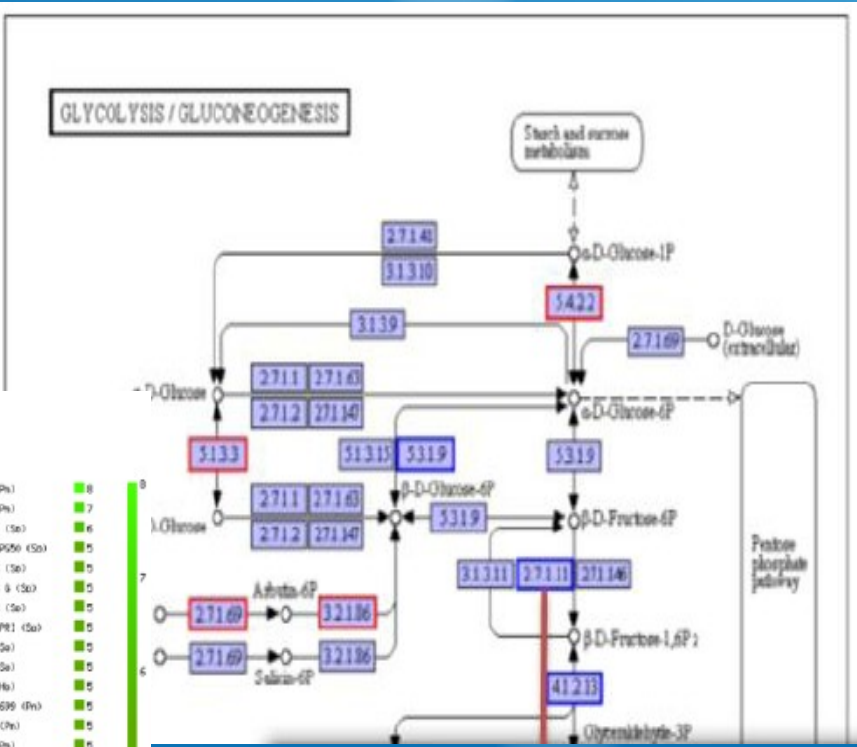
BIOINFORMATIQUE





the orientation / Co-occurrence

HE001	HE002	HE003	HE004	He
HE005	HE006	HE007	HE008	He
HE009	HE010	HE011	HE012	He
HE013	HE014	HE015	HE016	He
HE017	HE018	HE019	HE020	He
HE021	HE022	HE023	HE024	He
HE025	HE026	HE027	HE028	He
HE029	HE030	HE031	HE032	He
HE033	HE034	HE035	HE036	He
HE037	HE038	HE039	HE040	He
HE041	HE042	HE043	HE044	He
HE045	HE046	HE047	HE048	He
HE049	HE050	HE051	HE052	He
HE053	HE054	HE055	HE056	He
HE057	HE058	HE059	HE060	He
HE061	HE062	HE063	HE064	He
HE065	HE066	HE067	HE068	He
HE069	HE070	HE071	HE072	He
HE073	HE074	HE075	HE076	He
HE077	HE078	HE079	HE080	He
HE081	HE082	HE083	HE084	He
HE085	HE086	HE087	HE088	He
HE089	HE090	HE091	HE092	He
HE093	HE094	HE095	HE096	He
HE097	HE098	HE099	HE100	He
HE101	HE102	HE103	HE104	He
HE105	HE106	HE107	HE108	He
HE109	HE110	HE111	HE112	He
HE113	HE114	HE115	HE116	He
HE117	HE118	HE119	HE120	He
HE121	HE122	HE123	HE124	He
HE125	HE126	HE127	HE128	He
HE129	HE130	HE131	HE132	He
HE133	HE134	HE135	HE136	He
HE137	HE138	HE139	HE140	He
HE141	HE142	HE143	HE144	He
HE145	HE146	HE147	HE148	He
HE149	HE150	HE151	HE152	He
HE153	HE154	HE155	HE156	He
HE157	HE158	HE159	HE160	He
HE161	HE162	HE163	HE164	He
HE165	HE166	HE167	HE168	He
HE169	HE170	HE171	HE172	He
HE173	HE174	HE175	HE176	He
HE177	HE178	HE179	HE180	He
HE181	HE182	HE183	HE184	He
HE185	HE186	HE187	HE188	He
HE189	HE190	HE191	HE192	He
HE193	HE194	HE195	HE196	He
HE197	HE198	HE199	HE200	He
HE201	HE202	HE203	HE204	He
HE205	HE206	HE207	HE208	He
HE209	HE210	HE211	HE212	He
HE213	HE214	HE215	HE216	He
HE217	HE218	HE219	HE220	He
HE221	HE222	HE223	HE224	He
HE225	HE226	HE227	HE228	He
HE229	HE230	HE231	HE232	He
HE233	HE234	HE235	HE236	He
HE237	HE238	HE239	HE240	He
HE241	HE242	HE243	HE244	He
HE245	HE246	HE247	HE248	He
HE249	HE250	HE251	HE252	He
HE253	HE254	HE255	HE256	He
HE257	HE258	HE259	HE260	He
HE261	HE262	HE263	HE264	He
HE265	HE266	HE267	HE268	He
HE269	HE270	HE271	HE272	He
HE273	HE274	HE275	HE276	He
HE277	HE278	HE279	HE280	He
HE281	HE282	HE283	HE284	He
HE285	HE286	HE287	HE288	He
HE289	HE290	HE291	HE292	He
HE293	HE294	HE295	HE296	He
HE297	HE298	HE299	HE300	He
HE301	HE302	HE303	HE304	He
HE305	HE306	HE307	HE308	He
HE309	HE310	HE311	HE312	He
HE313	HE314	HE315	HE316	He
HE317	HE318	HE319	HE320	He
HE321	HE322	HE323	HE324	He
HE325	HE326	HE327	HE328	He
HE329	HE330	HE331	HE332	He
HE333	HE334	HE335	HE336	He
HE337	HE338	HE339	HE340	He
HE341	HE342	HE343	HE344	He
HE345	HE346	HE347	HE348	He
HE349	HE350	HE351	HE352	He
HE353	HE354	HE355	HE356	He
HE357	HE358	HE359	HE360	He
HE361	HE362	HE363	HE364	He
HE365	HE366	HE367	HE368	He
HE369	HE370	HE371	HE372	He
HE373	HE374	HE375	HE376	He
HE377	HE378	HE379	HE380	He
HE381	HE382	HE383	HE384	He
HE385	HE386	HE387	HE388	He
HE389	HE390	HE391	HE392	He
HE393	HE394	HE395	HE396	He
HE397	HE398	HE399	HE400	He
HE401	HE402	HE403	HE404	He
HE405	HE406	HE407	HE408	He
HE409	HE410	HE411	HE412	He
HE413	HE414	HE415	HE416	He
HE417	HE418	HE419	HE420	He
HE421	HE422	HE423	HE424	He
HE425	HE426	HE427	HE428	He
HE429	HE430	HE431	HE432	He
HE433	HE434	HE435	HE436	He
HE437	HE438	HE439	HE440	He
HE441	HE442	HE443	HE444	He
HE445	HE446	HE447	HE448	He
HE449	HE450	HE451	HE452	He
HE453	HE454	HE455	HE456	He
HE457	HE458	HE459	HE460	He
HE461	HE462	HE463	HE464	He
HE465	HE466	HE467	HE468	He
HE469	HE470	HE471	HE472	He
HE473	HE474	HE475	HE476	He
HE477	HE478	HE479	HE480	He
HE481	HE482	HE483	HE484	He
HE485	HE486	HE487	HE488	He
HE489	HE490	HE491	HE492	He
HE493	HE494	HE495	HE496	He
HE497	HE498	HE499	HE500	He
HE501	HE502	HE503	HE504	He
HE505	HE506	HE507	HE508	He
HE509	HE510	HE511	HE512	He
HE513	HE514	HE515	HE516	He
HE517	HE518	HE519	HE520	He
HE521	HE522	HE523	HE524	He
HE525	HE526	HE527	HE528	He
HE529	HE530	HE531	HE532	He
HE533	HE534	HE535	HE536	He
HE537	HE538	HE539	HE540	He
HE541	HE542	HE543	HE544	He
HE545	HE546	HE547	HE548	He
HE549	HE550	HE551	HE552	He
HE553	HE554	HE555	HE556	He
HE557	HE558	HE559	HE560	He
HE561	HE562	HE563	HE564	He
HE565	HE566	HE567	HE568	He
HE569	HE570	HE571	HE572	He
HE573	HE574	HE575	HE576	He
HE577	HE578	HE579	HE580	He
HE581	HE582	HE583	HE584	He
HE585	HE586	HE587	HE588	He
HE589	HE590	HE591	HE592	He
HE593	HE594	HE595	HE596	He
HE597	HE598	HE599	HE600	He
HE601	HE602	HE603	HE604	He
HE605	HE606	HE607	HE608	He
HE609	HE610	HE611	HE612	He
HE613	HE614	HE615	HE616	He
HE617	HE618	HE619	HE620	He
HE621	HE622	HE623	HE624	He
HE625	HE626	HE627	HE628	He
HE629	HE630	HE631	HE632	He
HE633	HE634	HE635	HE636	He
HE637	HE638	HE639	HE640	He
HE641	HE642	HE643	HE644	He
HE645	HE646	HE647	HE648	He
HE649	HE650	HE651	HE652	He
HE653	HE654	HE655	HE656	He
HE657	HE658	HE659	HE660	He
HE661	HE662	HE663	HE664	He
HE665	HE666	HE667	HE668	He
HE669	HE670	HE671	HE672	He
HE673	HE674	HE675	HE676	He
HE677	HE678	HE679	HE680	He
HE681	HE682	HE683	HE684	He
HE685	HE686	HE687	HE688	He
HE689	HE690	HE691	HE692	He
HE693	HE694	HE695	HE696	He
HE697	HE698	HE699	HE700	He
HE701	HE702	HE703	HE704	He
HE705	HE706	HE707	HE708	He
HE709	HE710	HE711	HE712	He
HE713	HE714	HE715	HE716	He
HE717	HE718	HE719	HE720	He
HE721	HE722	HE723	HE724	He
HE725	HE726	HE727	HE728	He
HE729	HE730	HE731	HE732	He
HE733	HE734	HE735	HE736	He
HE737	HE738	HE739	HE740	He
HE741	HE742	HE743	HE744	He
HE745	HE746	HE747	HE748	He
HE749	HE750	HE751	HE752	He
HE753	HE754	HE755	HE756	He
HE757	HE758	HE759	HE760	He
HE761	HE762	HE763	HE764	He
HE765	HE766	HE767	HE768	He
HE769	HE770	HE771	HE772	He
HE773	HE774	HE775	HE776	He
HE777	HE778	HE779	HE780	He
HE781	HE782	HE783	HE784	He
HE785	HE786	HE787	HE788	He
HE789	HE790	HE791	HE792	He
HE793	HE794	HE795	HE796	He
HE797	HE798	HE799	HE800	He
HE801	HE802	HE803	HE804	He
HE805	HE806	HE807	HE808	He
HE809	HE810	HE811	HE812	He
HE813	HE814	HE815	HE816	He
HE817	HE818	HE819	HE820	He
HE821	HE822	HE823	HE824	He
HE825	HE826	HE827	HE828	He
HE829	HE830	HE831	HE832	He
HE833	HE834	HE835	HE836	He
HE837	HE838	HE839	HE840	He
HE841	HE842	HE843	HE844	He
HE845	HE846	HE847	HE848	He
HE849	HE850	HE851	HE852	He
HE853	HE854	HE855	HE856	He
HE857	HE858	HE859	HE860	He
HE861	HE862	HE863	HE864	He
HE865	HE866	HE867	HE868	He
HE869	HE870	HE871	HE872	He
HE873	HE874	HE875	HE876	He
HE877	HE878	HE879	HE880	He
HE881	HE882	HE883	HE884	He
HE885	HE886	HE887	HE888	He
HE889	HE890	HE891	HE892	He
HE893	HE894	HE895	HE896	He
HE897	HE898	HE899	HE900	He
HE901	HE902	HE903	HE904	He
HE905	HE906	HE907	HE908	He
HE909	HE910	HE911	HE912	He
HE913	HE914	HE915	HE916	He
HE917	HE918	HE919	HE920	He
HE921	HE922	HE923	HE924	He
HE925	HE926	HE927	HE928	He
HE929	HE930	HE931	HE932	He
HE933	HE934	HE935	HE936	He
HE937	HE938	HE939	HE940	He
HE941	HE942	HE943	HE944	He
HE945	HE946	HE947	HE948	He
HE949	HE950	HE951	HE952	He
HE953	HE954	HE955	HE956	He
HE957	HE958	HE959	HE960	He
HE961	HE962	HE963	HE964	He
HE965	HE966	HE967	HE968	He
HE969	HE970	HE971	HE972	He
HE973	HE974	HE975	HE976	He
HE977	HE978	HE979	HE980	He
HE981	HE982	HE983	HE984	He
HE985	HE986	HE987	HE988	He
HE989	HE990	HE991	HE992	He
HE993	HE994	HE995	HE996	He
HE997	HE998	HE999	HE1000	He





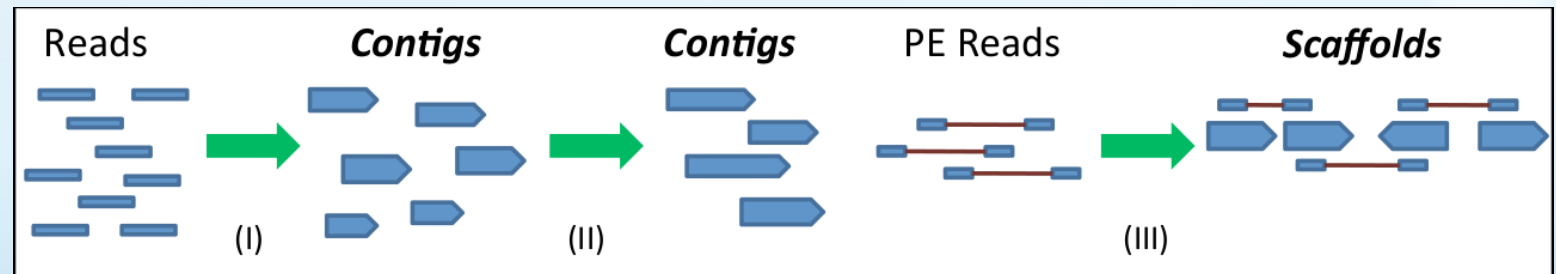
Evolmyco

- ANR evolmyco 2008–2011 : NGS sequencing of 10 new species or strains
- Unfinished genomes produced, with a large number of contigs
- Evident problem in assemblies in regard of existing genomes in MolliGen
- Results not usable as such → finishing strategy



Finishing of genome assemblies

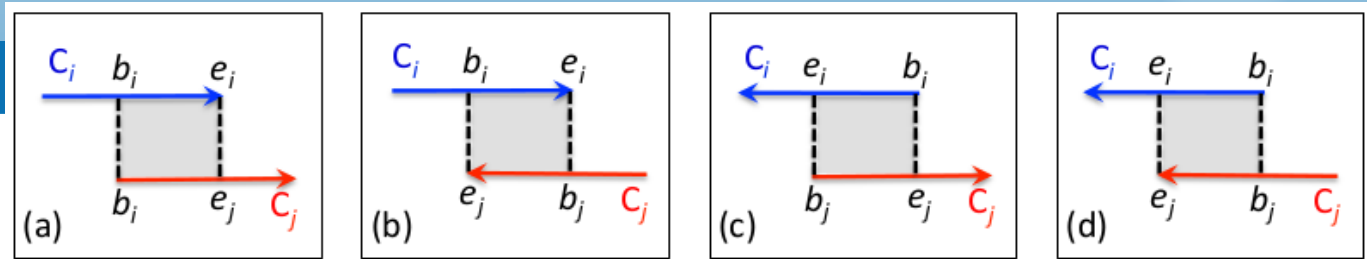
Genome assembly overview



- **Goals :**
 - Fill the gaps and **extend contigs**
 - Targeted sequencing to resolve misassemblies and sequence gaps
- **In silico finishing**
 - Achieve a better quality assembly (w.r.t. fragmentation)
 - Typically by assembly combination

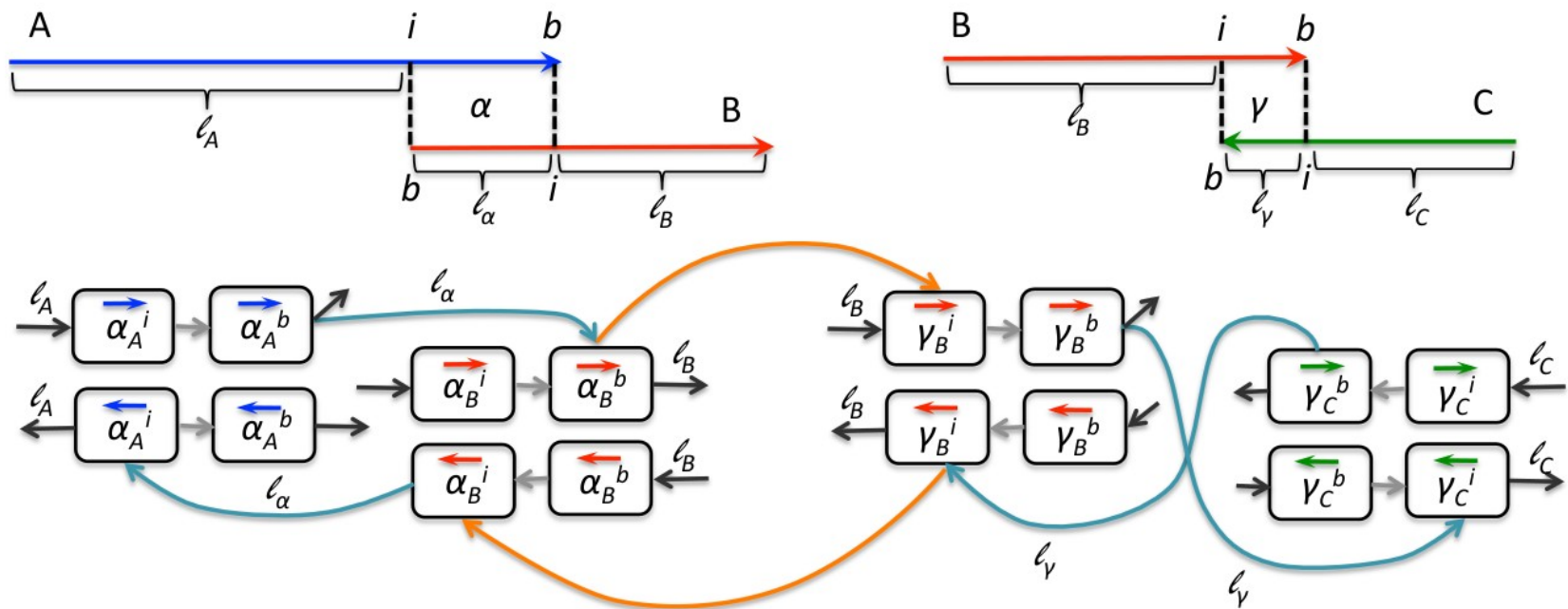
How we *Mix* assemblies

- 2 or more input assemblies - No reference genome, no additional raw reads, no « privileged » assembly
- *Mix* : Input data
 - Assemblies $\mathbb{A}_i = \{C_i\}$ pooled together $\mathbb{A} = \cup \mathbb{A}_i$
 - Alignments set \mathcal{A} : each $\mathbb{A}_i \in \mathbb{A}$ aligned against the others
- 4 nodes to represent an alignment zone on a contig
 - Limits of the alignment zone : b (boundary) and i (internal)
 - How the alignment zone is read : \rightarrow (forward) and \leftarrow (reverse)



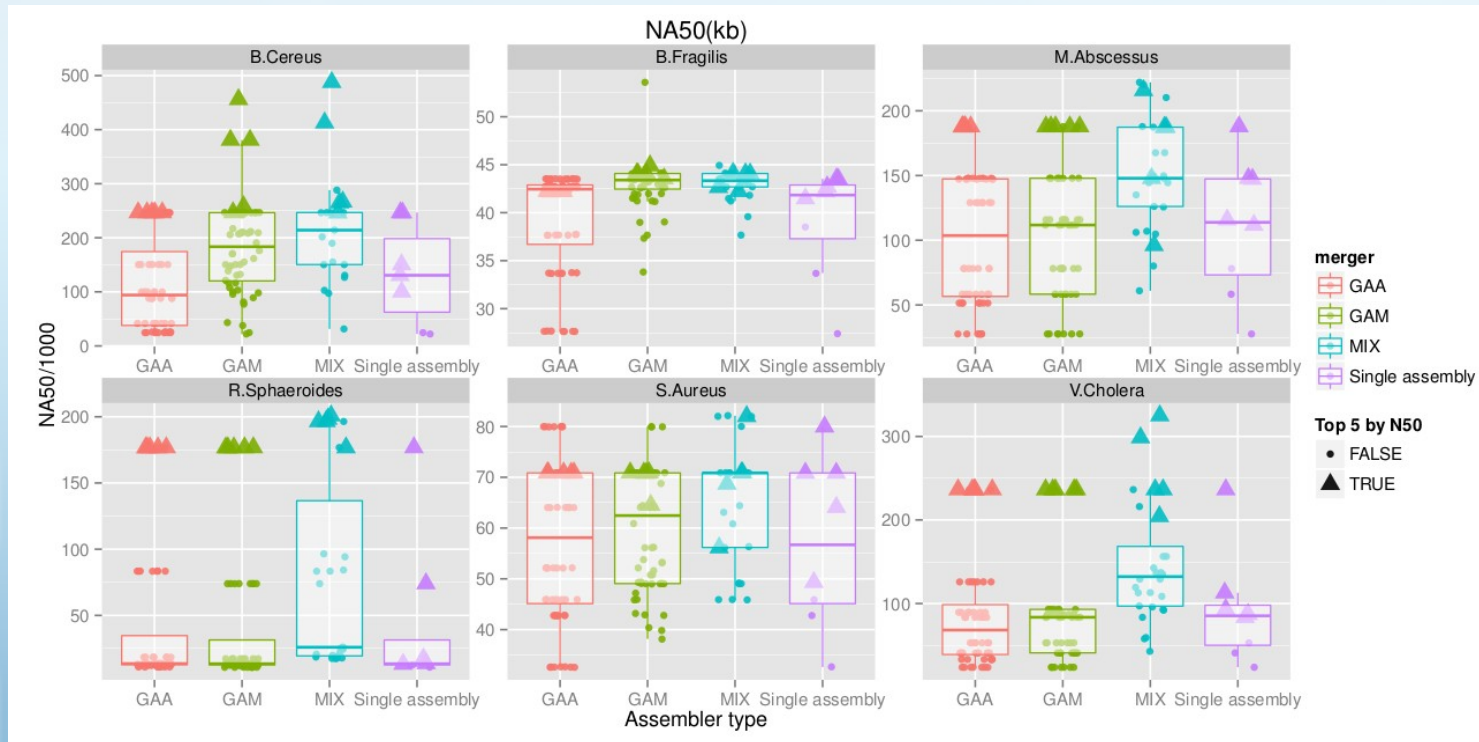
Extension graph

- What is *Mixed*?
 - Create an extension graph
 - Look for paths that maximize the path length



Performance evaluation : the GAGE-B benchmark (Magoc *et al.*, 2013)

- Compares 8 assemblers and 2 assembly mergers



→ Fully correlated with N50

Application to the *Mycoplasma*

- 10 *Mycoplasma* genomes
- Three assemblers (CLC, ABySS, MIRA)
- 454 (130~170 nt) and Illumina (36 nt)
- Different assemblers correctly assemble different regions
- *Mix* systematically reduces genome fragmentation
- Validation with core genome
- Annotation and submission to the MolliGen database

<https://github.com/cbib/MIX>



... next steps

- Resulting assemblies are integrated in MolliGen
- Comparison with existing genomes : core gene set & evolution scenario
- MIX will be used as a service in MolliGen to integrate new genomes directly from reads files



Thanks to

From CBiB

- Florence MAURIER
- Hayssam SOUEIDAN
- David BENABEN
- Macha NIKOLSKI

